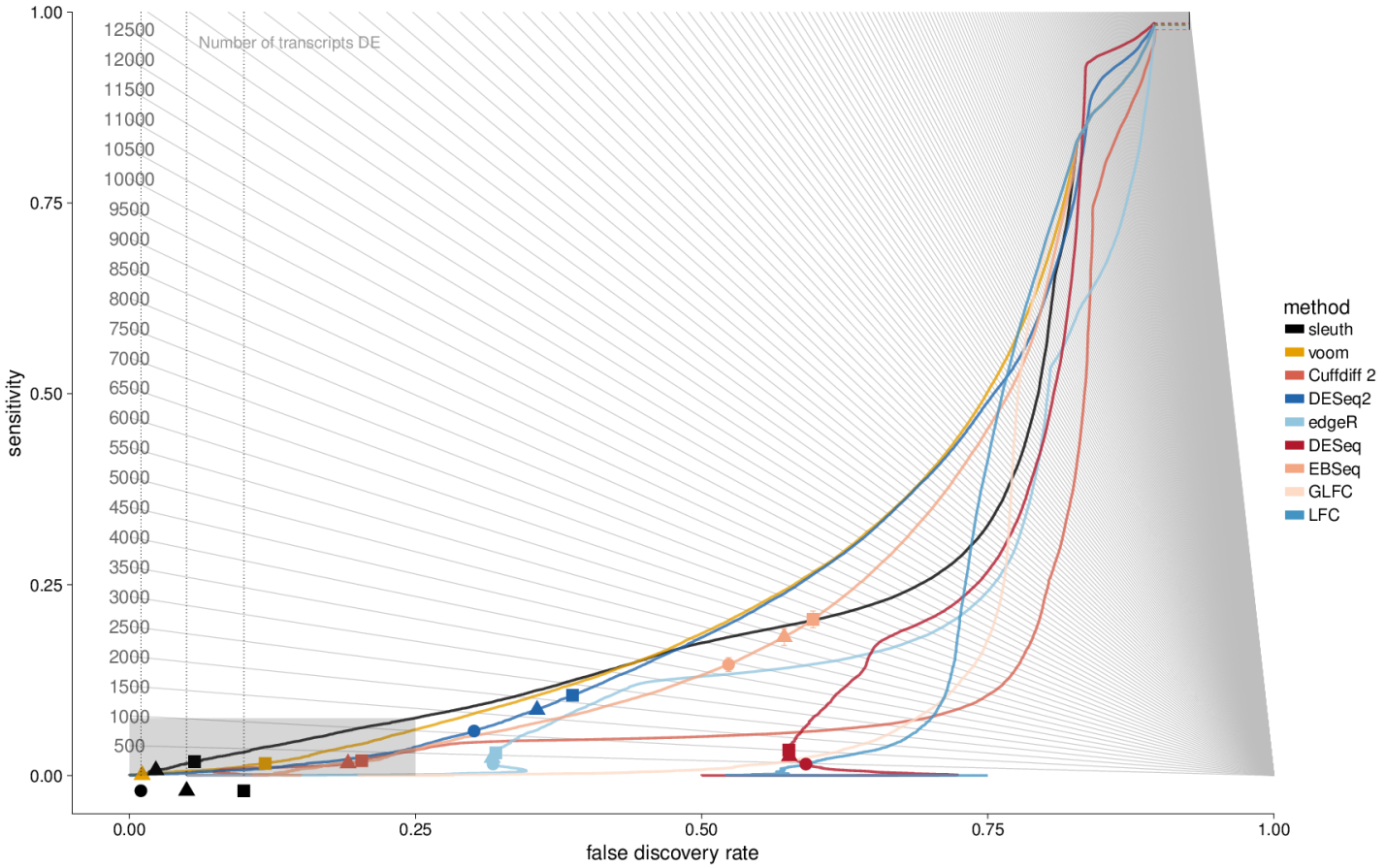
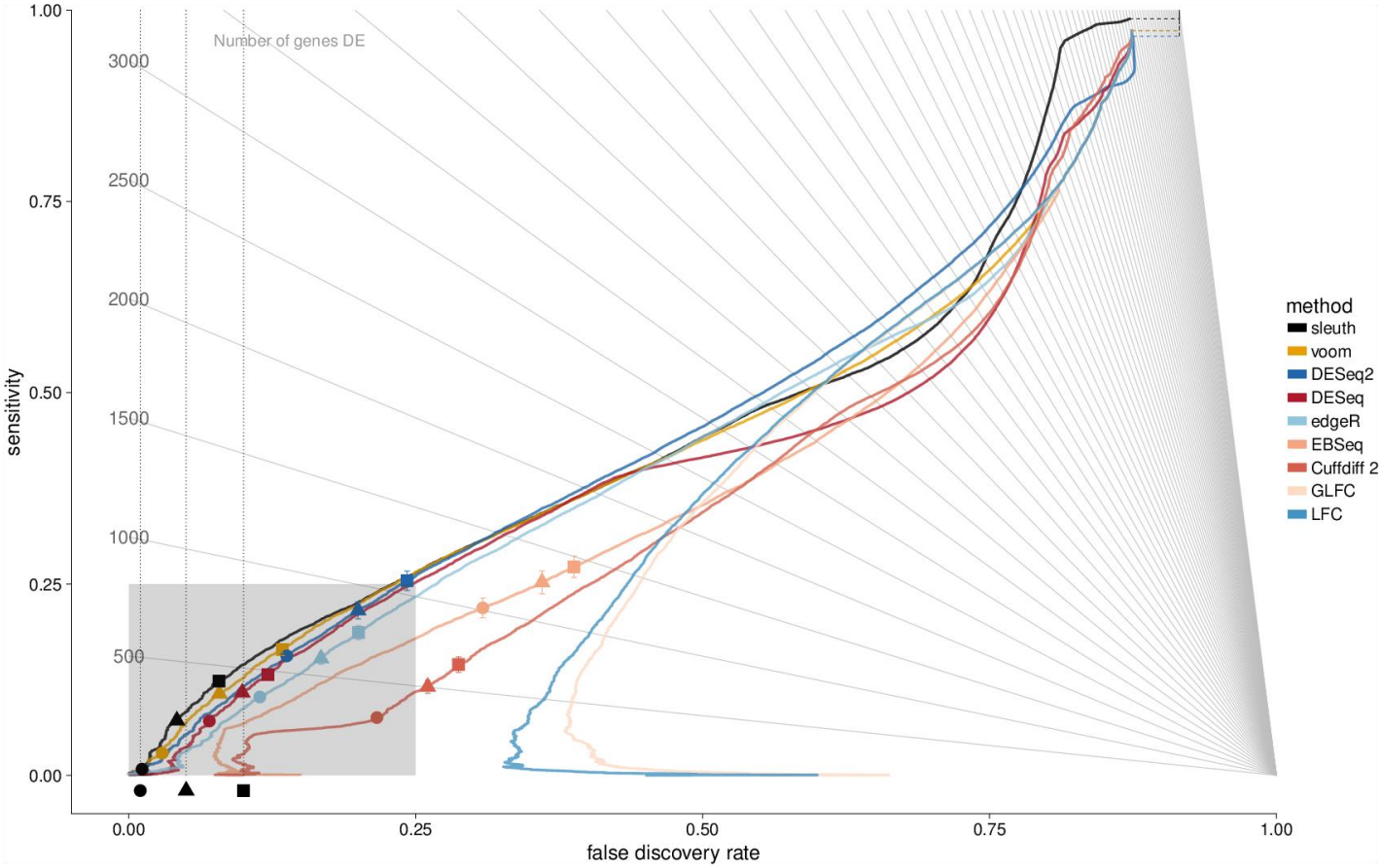


Supplementary Figure 1

Sensitivity versus FDR in the "effect from experiment" simulation at the transcript level.

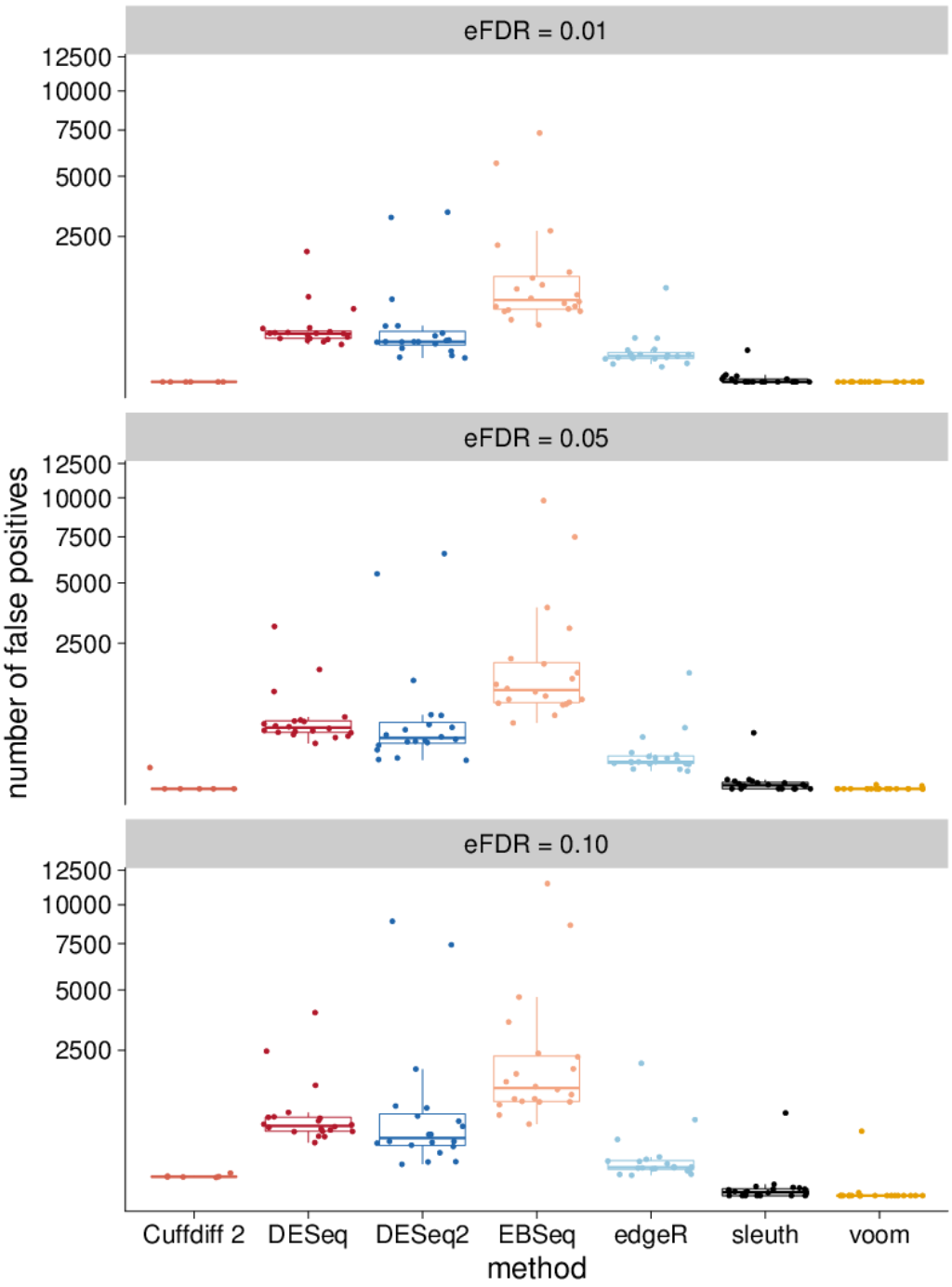
Zoomed out version of performance on effect from experiment simulation at the isoform level. The gray box in the bottom left-hand corner represents the zoomed in region in Figure 2. See Figure 2 caption for more information.



Supplementary Figure 2

Sensitivity versus FDR in the "effect from experiment" simulation at the gene level.

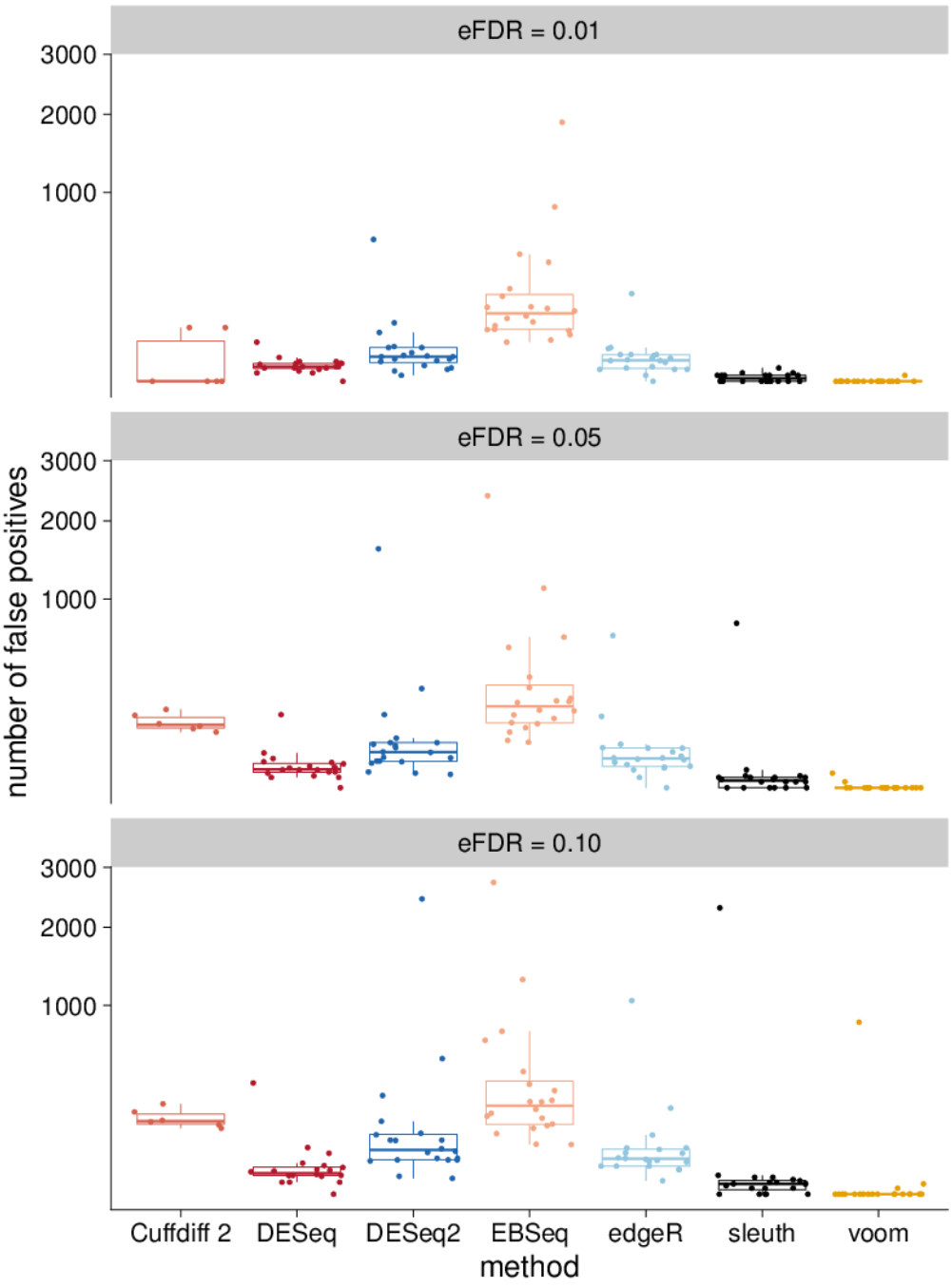
Zoomed out version of performance on effect from experiment simulation at the gene level. The gray box in the bottom left-hand corner represents the zoomed in region in Figure 2. See Figure 2 caption for more information.



Supplementary Figure 3

Null resampling experiment at the isoform level.

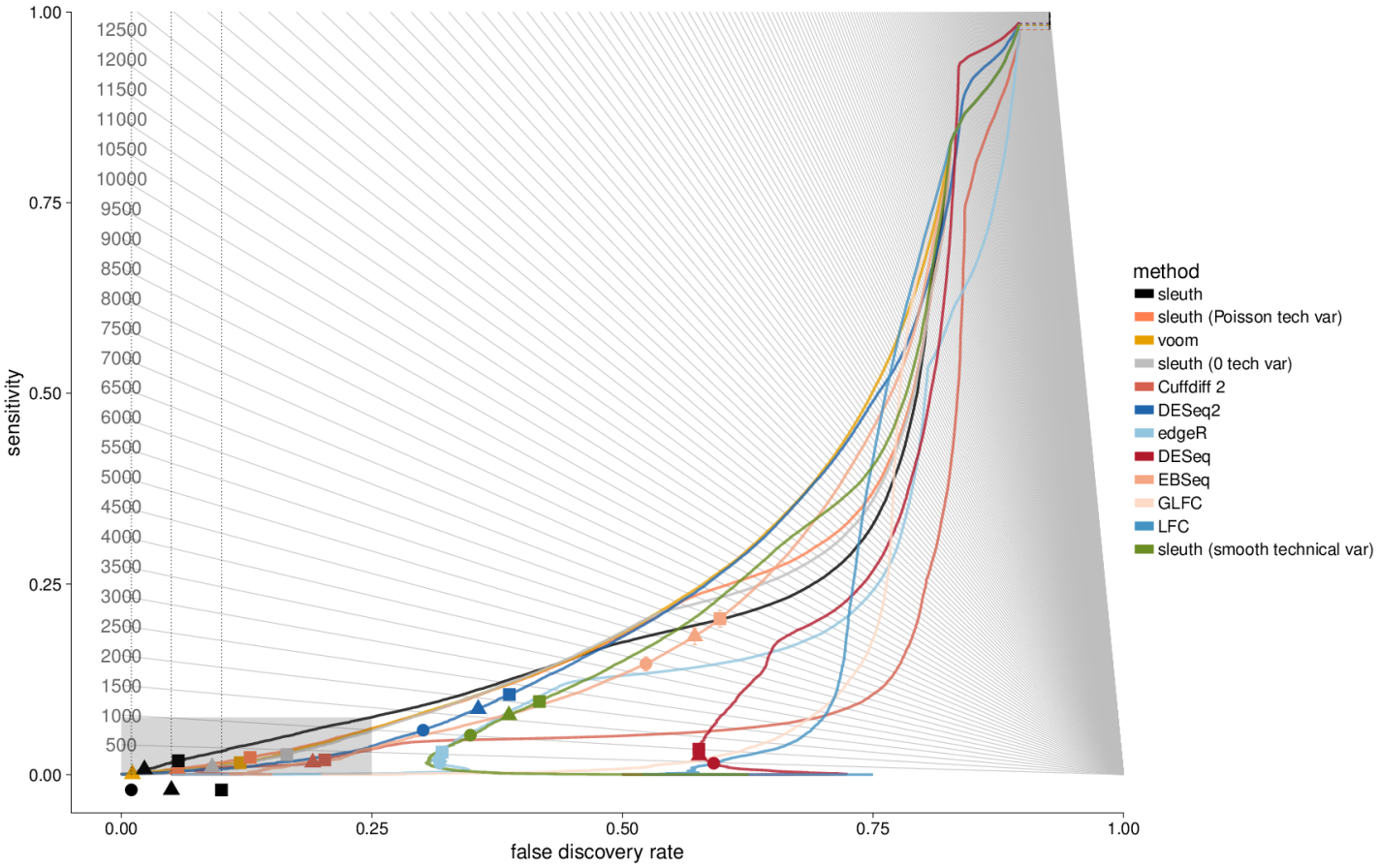
Each dot represents the number of false positives in a particular shuffling, and the box plot represents the distribution. Each point represents the number of false positives of a method on a single experiment. Each box plot contains hinges at the 25th and 75th percentile, a line at the median, and whiskers extending to the smallest/largest value no less/more than 1.5*IQR from the median.



Supplementary Figure 4

Null resampling experiment at the gene level.

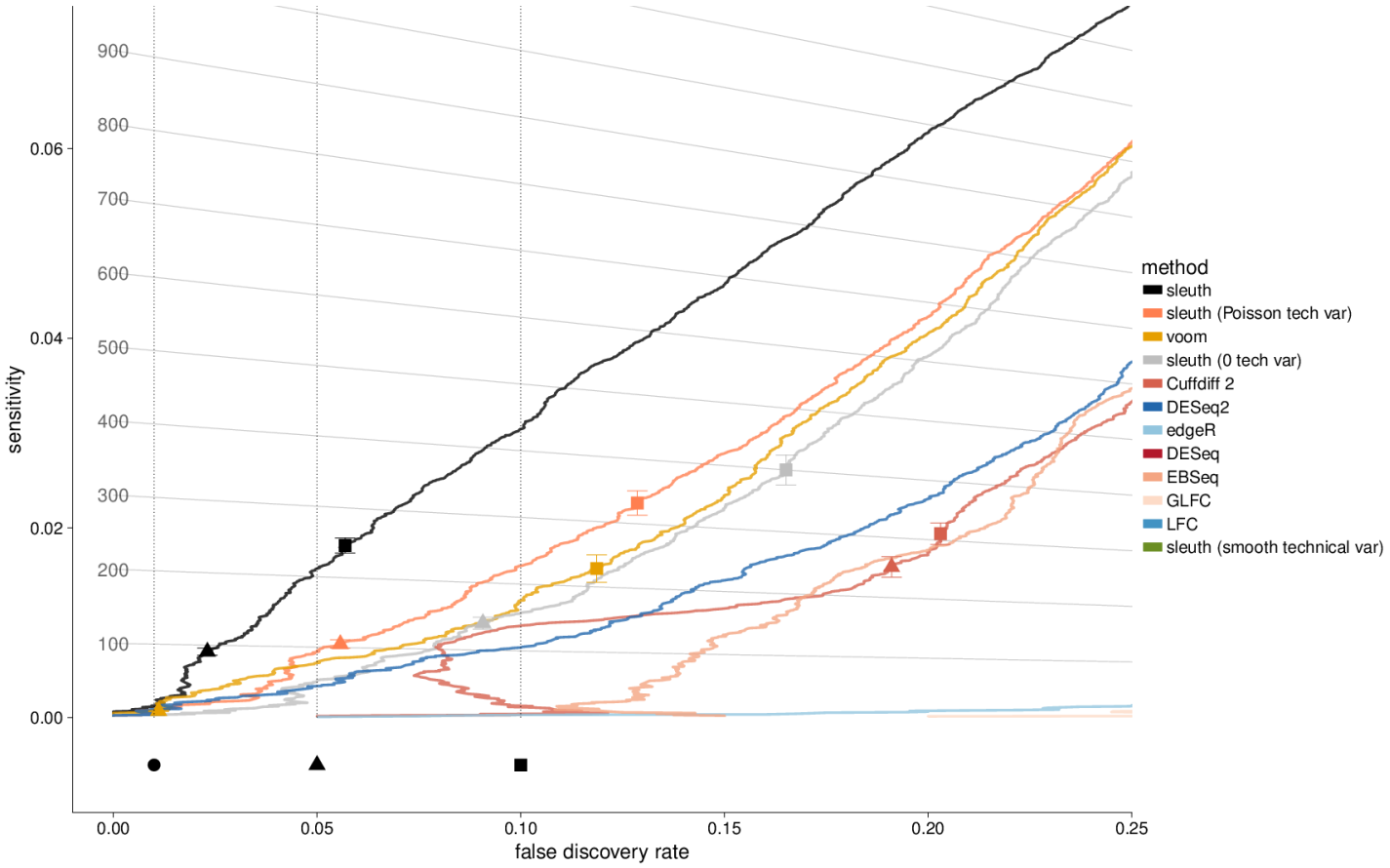
Each dot represents the number of false positives in a particular shuffling, and the box plot represents the distribution. Each point represents the number of false positives of a method on a single experiment. Each box plot contains hinges at the 25th and 75th percentile, a line at the median, and whiskers extending to the smallest/largest value no less/more than 1.5*IQR from the median.



Supplementary Figure 5

Sensitivity versus FDR in the "effect from experiment" simulation at the transcript level including alternative variance estimators.

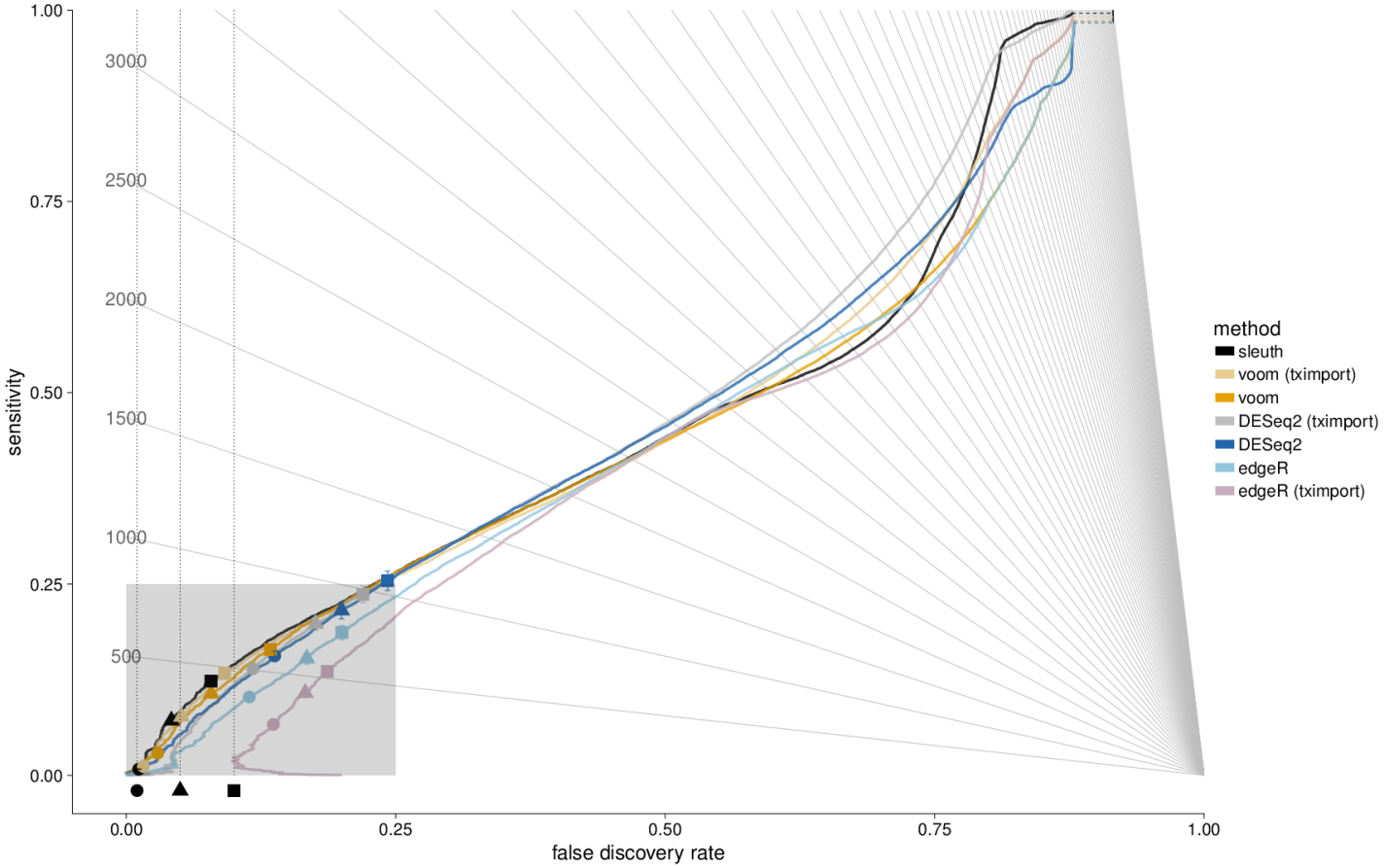
Zoomed out version of performance on effect from reference simulation at the isoform level introducing additional variance estimators for sleuth. The gray box in the bottom left-hand corner represents the zoomed in region in Supplementary Figure 6.



Supplementary Figure 6

Sensitivity versus FDR in the "effect from experiment" simulation at the transcript level including alternative variance estimators.

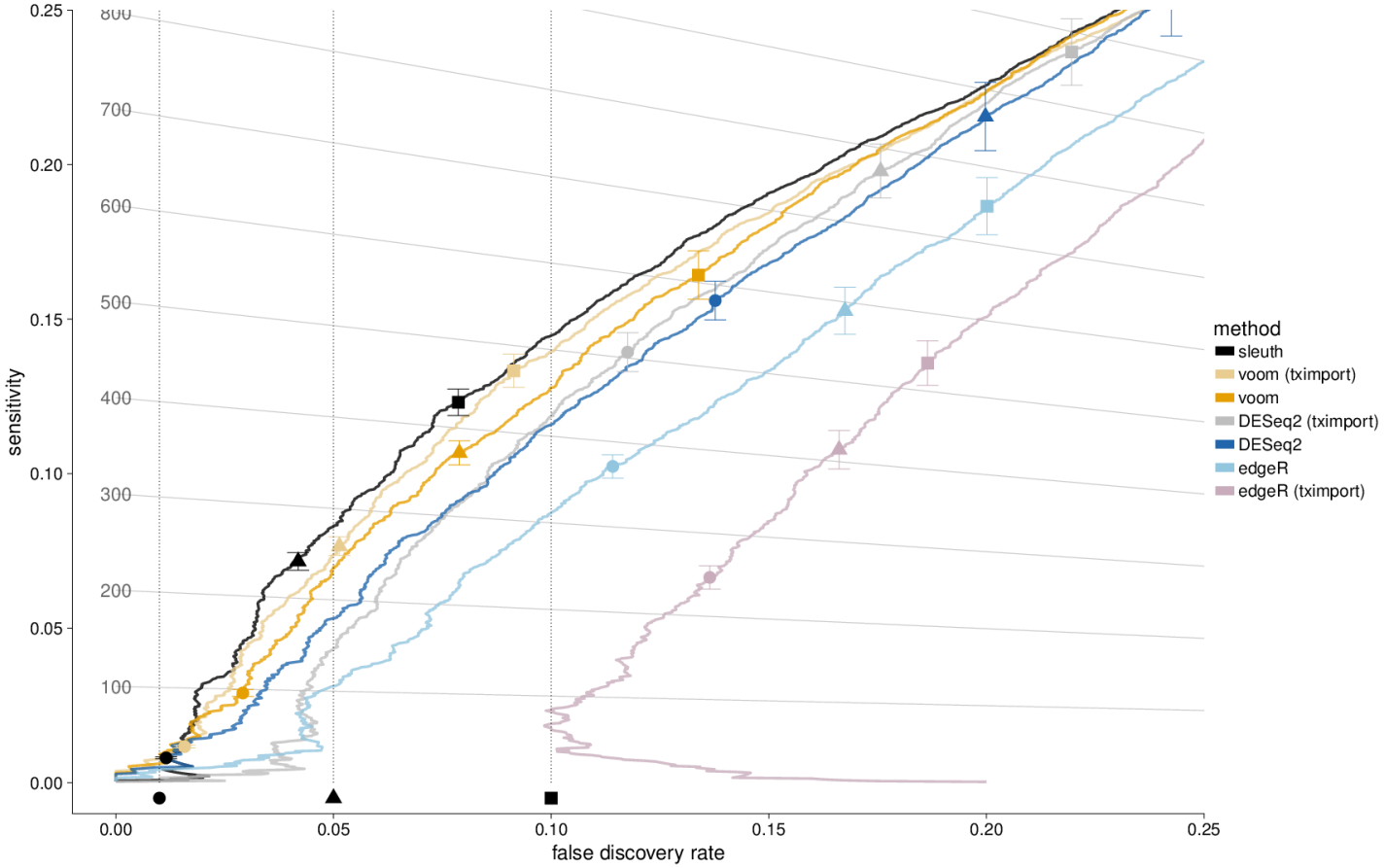
Zoomed in version of performance on effect from reference simulation at the isoform level introducing additional variance estimators for sleuth.



Supplementary Figure 7

Sensitivity versus FDR in the "effect from experiment" simulation at the gene level including tximport.

Zoomed out version of performance on effect from reference simulation at the gene level substituting tximport for featureCounts. The gray box in the bottom left-hand corner represents the zoomed in region in Supplementary Figure 8.



Supplementary Figure 8

Sensitivity versus FDR in the "effect from experiment" simulation at the gene level including tximport.

Zoomed in version of performance on effect from reference simulation at the gene level substituting tximport for featureCounts.

Supplementary Note 1: Details of the simulations

Harold Pimentel Nicolas Bray Suzette Puente Páll Melsted
Lior Pachter

April 27, 2017

1 Description of filters of benchmarked programs

In this section we describe the filtering procedures of the programs benchmarked (in a few cases methods did not specify filtering procedures so we selected one for them). Supplementary Table 1 shows the filters used at the transcript and gene level:

method	**isoform mode filter**	**gene mode filter**
Cuffdiff 2	Cuffdiff 2	Cuffdiff 2
DESeq	DESeq	DESeq
DESeq2	DESeq2	DESeq2
edgeR	edgeR	edgeR
EBSeq	sleuth	edgeR
GLFC	sleuth	edgeR
LFC	sleuth	edgeR
sleuth	sleuth	sleuth
voom	sleuth	edgeR

Supplementary Table 1: The filters used with each program.

1.1 Cuffdiff 2

The default filter for the program is described in the Cuffdiff2 manual as: "The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples. If no testing is performed, changes in the locus are deemed not significance, and the locus' observed changes don't contribute to correction for multiple testing. The default is 10 fragment alignments."

1.2 DESeq

The DESeq vignette describes a filter discarding the lowest 40% of expressed features, where expression is defined as the total number of counts across all experiments. In some cases more than 40% of the features were lowly expressed so we implemented a slightly modified

Nature Methods: doi:10.1038/nmeth.4324

version that first applied the DESeq2 filter.

```
DESeq_filter <- function(mat, ...) {
  # a modified version of the DESeq filter to first remove things that are 0
  # before doing the quantile filter
  nonzero <- DESeq2_filter(mat)
  rs <- rowSums(mat[nonzero, ])
  theta <- 0.4
  use <- (rs > quantile(rs, probs=theta))
  ret <- nonzero
  ret[nonzero] <- use
  ret
}
```

1.3 DESeq2

The filter is described in the DESeq2 vignette. It removes features whose total counts across all experiments is less than 2:

```
DESeq2_filter <- function(mat, ...) {
  rowSums(mat) > 1
}
```

1.4 edgeR

The filter is described in the edgeR vignette. It removes features where less than 2 experiments contain less than or equal to 1 count per million:

```
edgeR_filter <- function(mat, ...) {
  rowSums(cpm(mat) > 1) >= 2
}
```

1.5 EBSeq

Based on the EBSeq vignette we decided to use the sleuth filter at the isoform level and edgeR filter at the gene level.

1.6 voom

The voom vignette states "The limma-voom method assumes that rows with zero or very low counts have been removed", so we decided to use the sleuth filter at the isoform level and edgeR at the gene level.

Nature Methods: doi:10.1038/nmeth.4324

2 Log fold change

We also benchmarked the approach of using log fold change between conditions to directly identify differentially expressed genes. Two alternatives are examined:

- LFC - log-fold change
- GLFC - geometric log-fold change

The definitions are as follows: Let A and B be two sets which contain samples from two conditions. The LFC for feature for a transcript or gene t is defined as:

$$\text{LFC}_t = \log\left(\frac{\text{mean}_{i\in B}(c_{ti}/\hat{s}_i)}{\text{mean}_{j\in A}(c_{tj}/\hat{s}_j)}\right).$$

Following [3], the GLFC for feature t is defined as:

$$\text{GLFC}_t = \text{mean}_{i\in B}\left(\log(c_{ti}/\hat{s}_i + 0.5)\right) - \text{mean}_{j\in A}\left(\log(c_{tj}/\hat{s}_j + 0.5)\right).$$

At the transcript level, kallisto counts were used for c_{ti}. At the gene level featureCounts were used for c_{ti}. DESeq2 normalization was performed on the raw counts for both methods.

3 Simulation details

3.1 Overview of simulations

The null distribution was learned by fitting negative binomials to kallisto transcript quantifications using the Cox-Reid estimator implemented in DESeq2 on the female Finnish population in the GEUVAIDS data [2]. While the model was still fit to lowly expressed transcripts, filtered transcripts were marked so that while reads would be generated, they would never be simulated as DE. Transcripts with low expression where the dispersion parameter could not be estimated had their dispersion set to the median dispersion. Minimum dispersion was set to 1e-6 so that reads could be simulated from the negative binomial distribution.

A model for DE simulation at the transcript level is then:

$$X_{tj} \sim \text{NegativeBinomial}(s_j\mu_t f_{tj}, \phi_t)$$

where t is the transcript and j is the sample. X_{tj} are the counts of transcript t in condition j, s_j is the sample specific size factor, μ_t is the mean number of counts learned from the experiment, f_{tj} is the transcript and sample specific fold change ($f_{tj} = 1$ if not differentially expressed) and ϕ_t is the transcript dispersion parameter. The effect size, f_{tj} is the same amongst the experimental conditions. For example, $f_{t1} = f_{t2} = f_{t3}$ and $f_{t4} = f_{t5} = f_{t6}$, but $f_{t1} \neq f_{t6}$ if transcript t is differentially expressed.

For the exact number of reads generated in each simulation, please see the tables at the end of this note.

3.2 Independent effect

20% isoforms were chosen at random to be differentially expressed. Log fold change was generated from a normal distribution truncated at 0, with mean 1.5 and standard deviation 1.

3.3 Correlated effect

20% of genes were chosen at random to be differentially expressed. If a gene was chosen differentially expressed, then its log fold change was chosen the same way as above for every isoform, except the direction was chosen to be the same for every single isoform in that gene.

3.4 Effect from experiment

Isoform level differential expression was performed using DESeq2 and sleuth on the Trapnell et al. data. Every gene that contained a differentially expressed isoform was considered differentially expressed. The union of genes found by DESeq2 and sleuth was noted, along with the number of isoforms, their fold change, and their rank and expression among other isoforms in that gene.

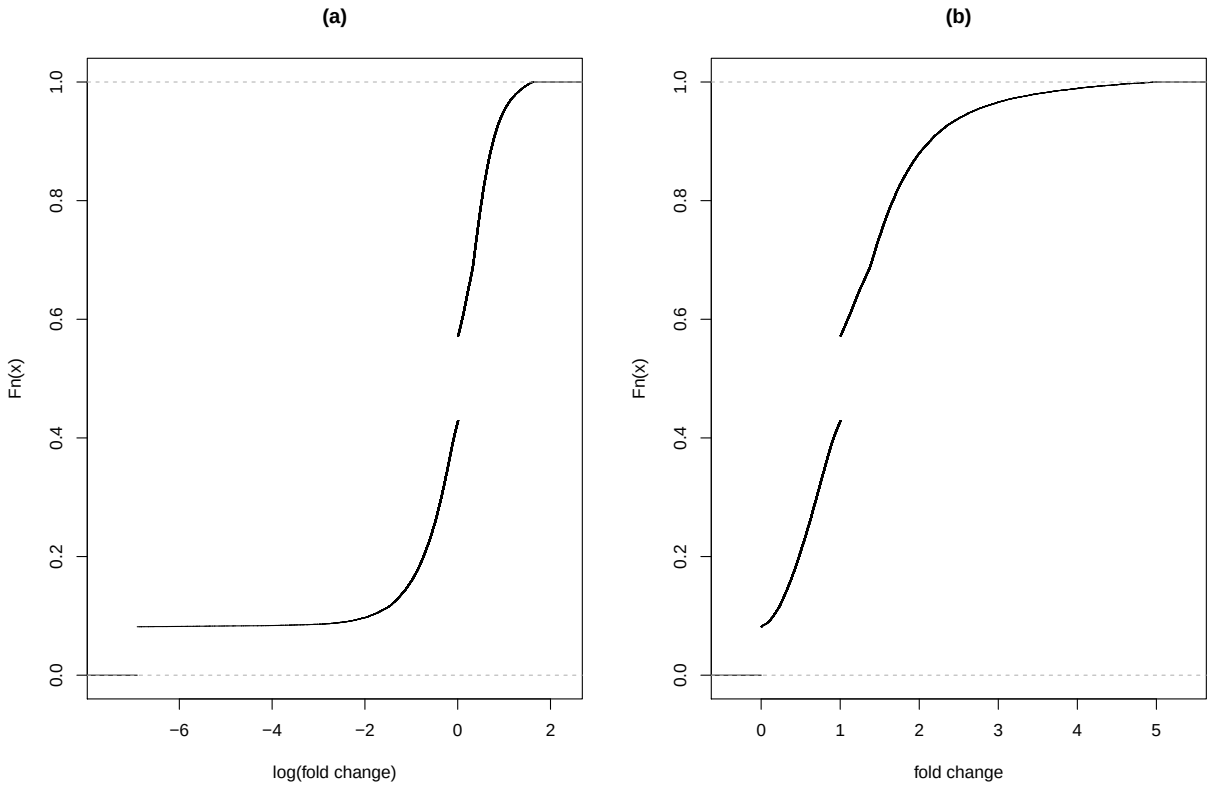
For fold change larger than 5, we generated a fold change from a truncated normal centered around the 75th percentile of fold change observed. Fold change was set to a minimum of 0.001.

To simulate gene level differential expression, 20% of genes were selected at random. A set of gene level fold change was chosen at random from genes with the same number of isoforms. The fold change for each isoform in that gene was then chosen to match the rank of expression in the experiment. This was done to prevent highly expressed isoforms from getting extremely large fold changes that might result from lowly expressed isoforms.

3.5 Learning effect sizes from an experiment

A distribution on fold changes differentially expressed transcripts and genes in the Trapnell *et al.* [4] experiment was estimated from the set of all genes found to be differentially expressed at FDR 0.05 by either DESeq2 or sleuth. While these programs were used to learn a distribution of fold change, in the simulation transcripts were chosen randomly for perturbation. Supplementary Figure SN1 shows the cumulative distribution of fold changes estimated from the data.

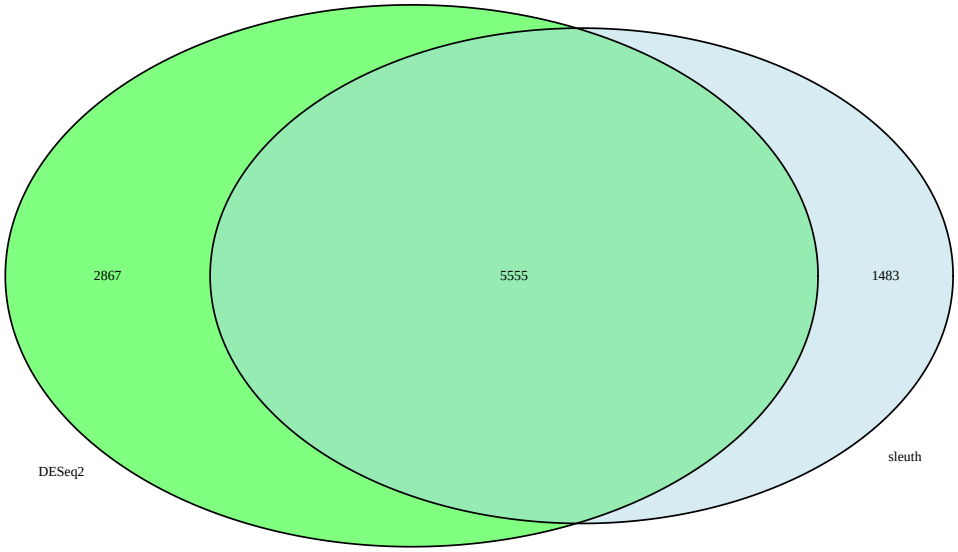
Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN1: Empirical cummulative distribution function (ECDF) of (a) log fold change and (b) fold change for differentially expressed isoforms.

The simulation was purposely designed to include perturbations of small effect, as these have been determined to be biologically important in a variety of settings [1]. However results were also stratified by effect size (see Supplementary Figure SN11) and for large effect sizes the large overlap between DESeq2 and sleuth predictions on the Trapnell *et al.* data ensured that the distribution was robust.

Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN2: Overlap of genes called differentially expressed using if at least one transcript in the gene was differentially expressed with sleuth, and using gene counting with DESeq2. These genes were used to learn the effect size from at the transcript level.

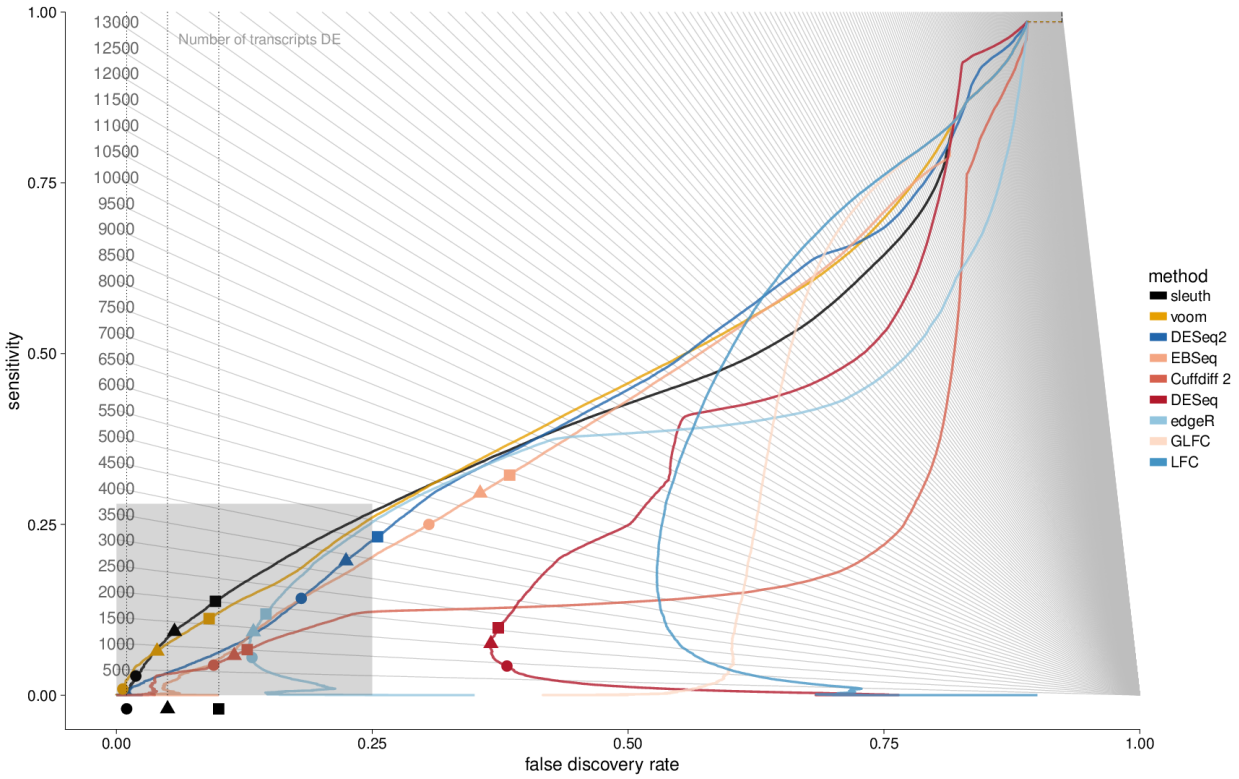
4 Performance on simulations

In this section we show how each of the methods benchmarked performs on the independent effect and correlated effect simulations with the filtering procedures described in Section 8. For context, we also show the performance on the effect from experiment simulation zoomed out to cover the entire spectrum of false discovery rate, and reparametrize it in terms of number of predictions and precision.

Nature Methods: doi:10.1038/nmeth.4324

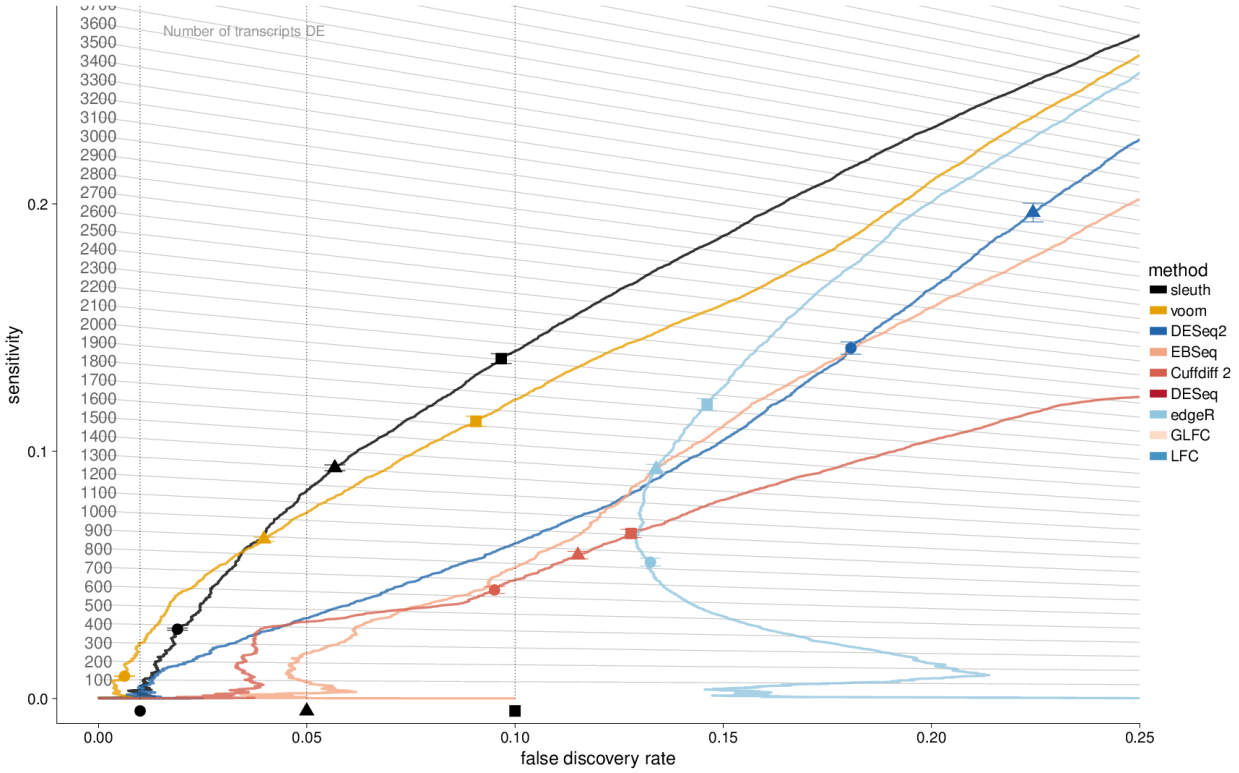
4.1 Independent effect simulation

4.1.1 Isoform level



Supplementary Figure SN3: Zoomed out version of performance on independent effect simulation at the isoform level.

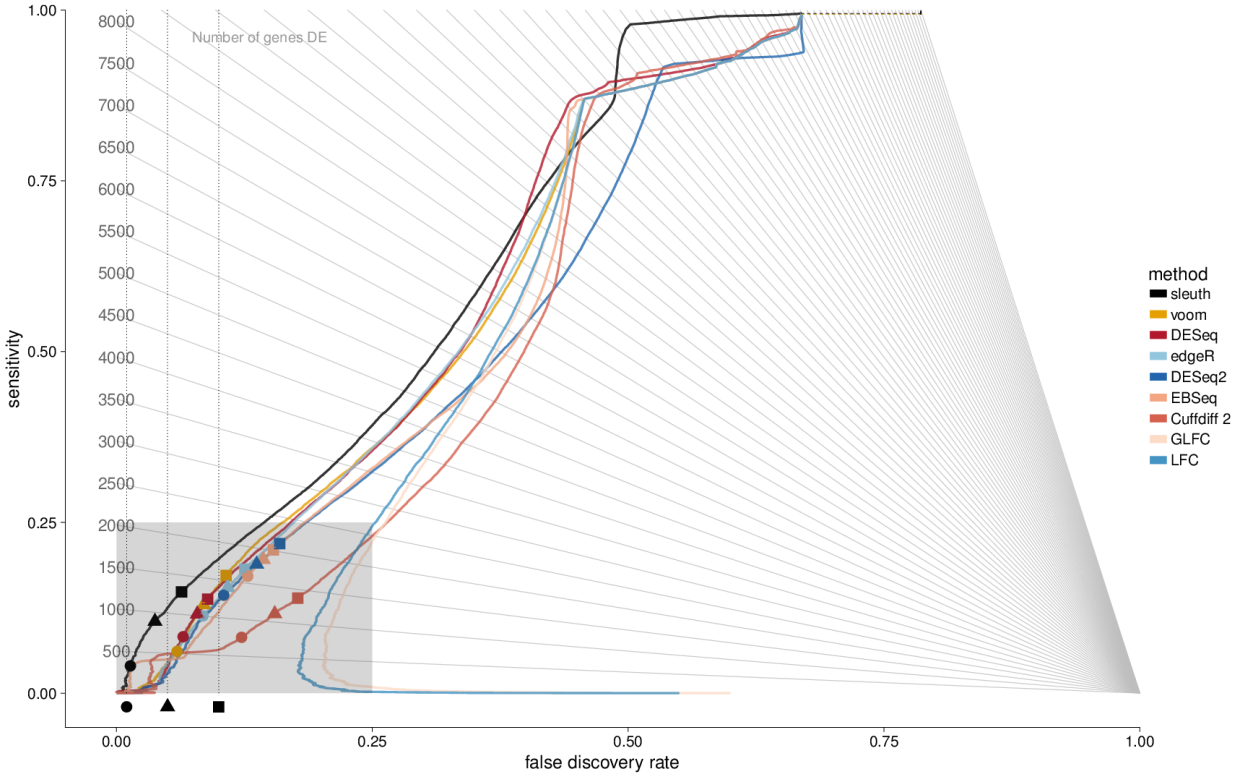
Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN4: Zoomed in version of performance on independent effect simulation at the isoform level.

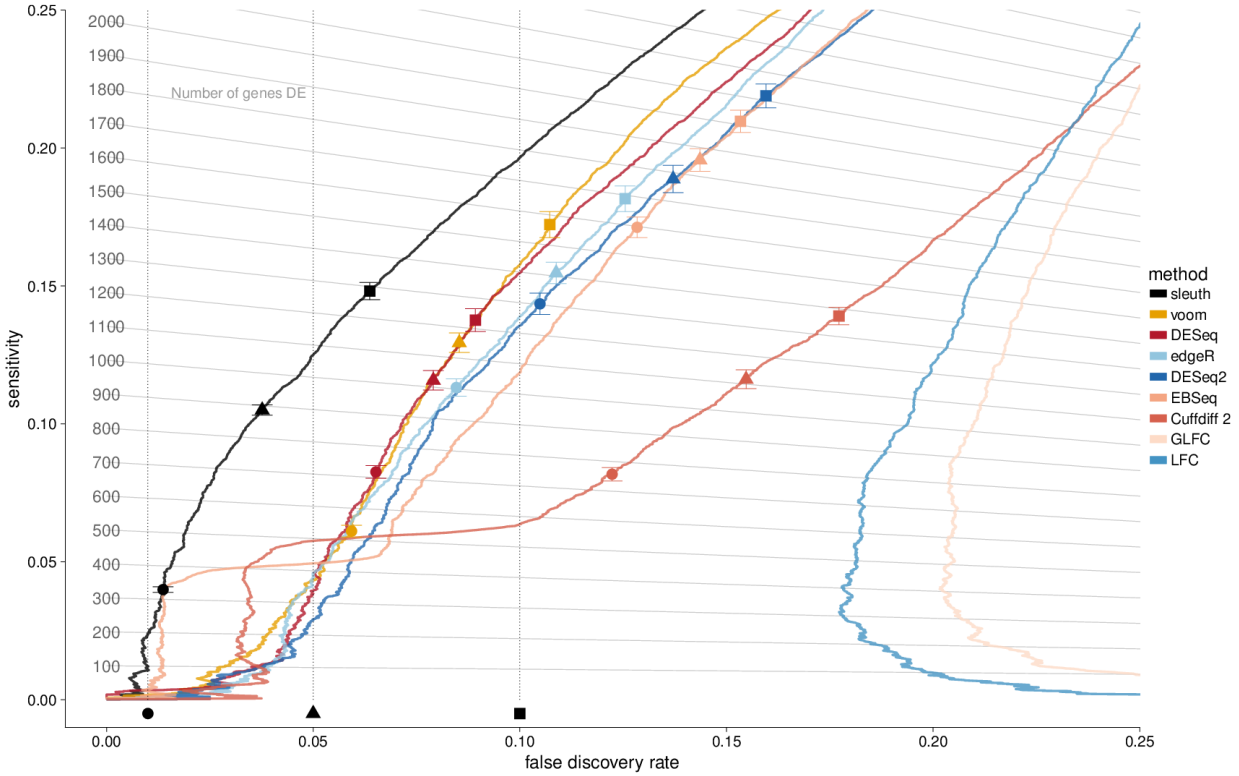
Nature Methods: doi:10.1038/nmeth.4324

4.1.2 Gene level



Supplementary Figure SN5: Zoomed out version of performance on independent effect simulation at the gene level.

Nature Methods: doi:10.1038/nmeth.4324

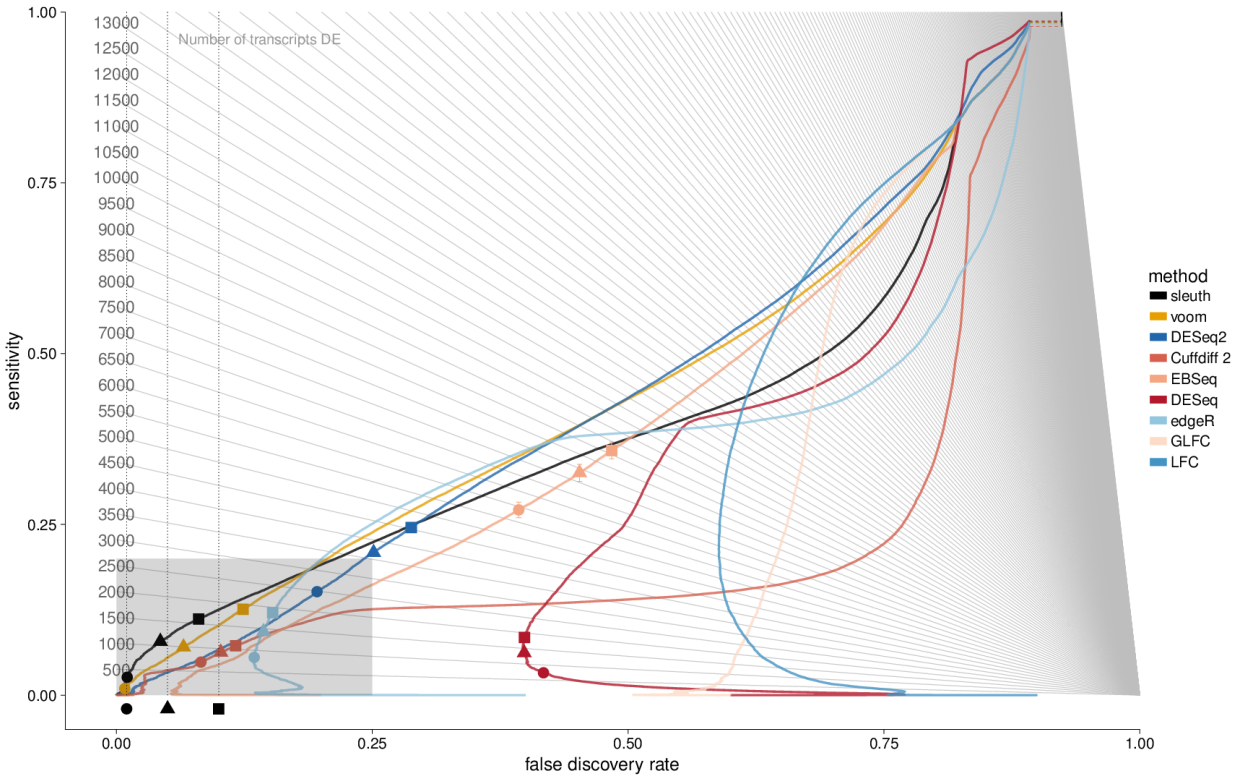


Supplementary Figure SN6: Zoomed in version of performance on independent effect simulation at the gene level.

Nature Methods: doi:10.1038/nmeth.4324

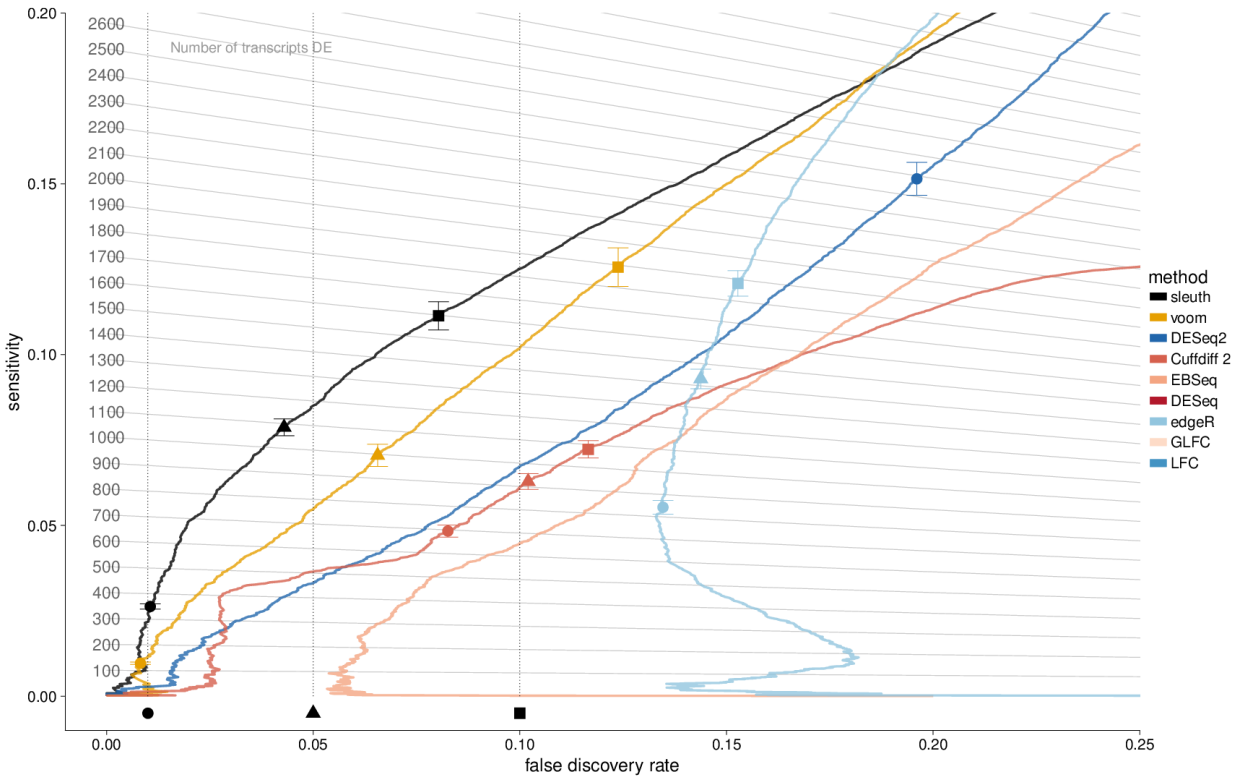
4.2 Correlated effect simulation

4.2.1 Isoform level



Supplementary Figure SN7: Zoomed out version of performance on correlated effect simulation at the isoform level.

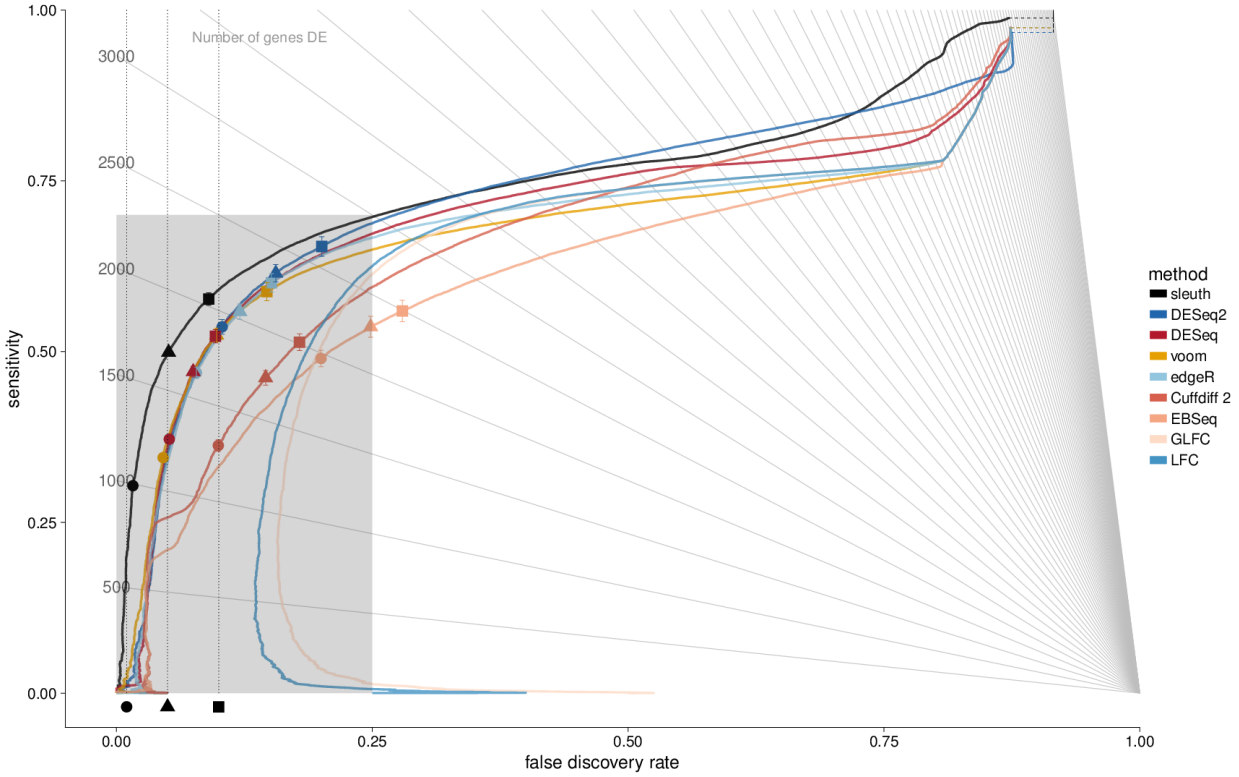
Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN8: Zoomed in version of performance on correlated effect simulation at the isoform level.

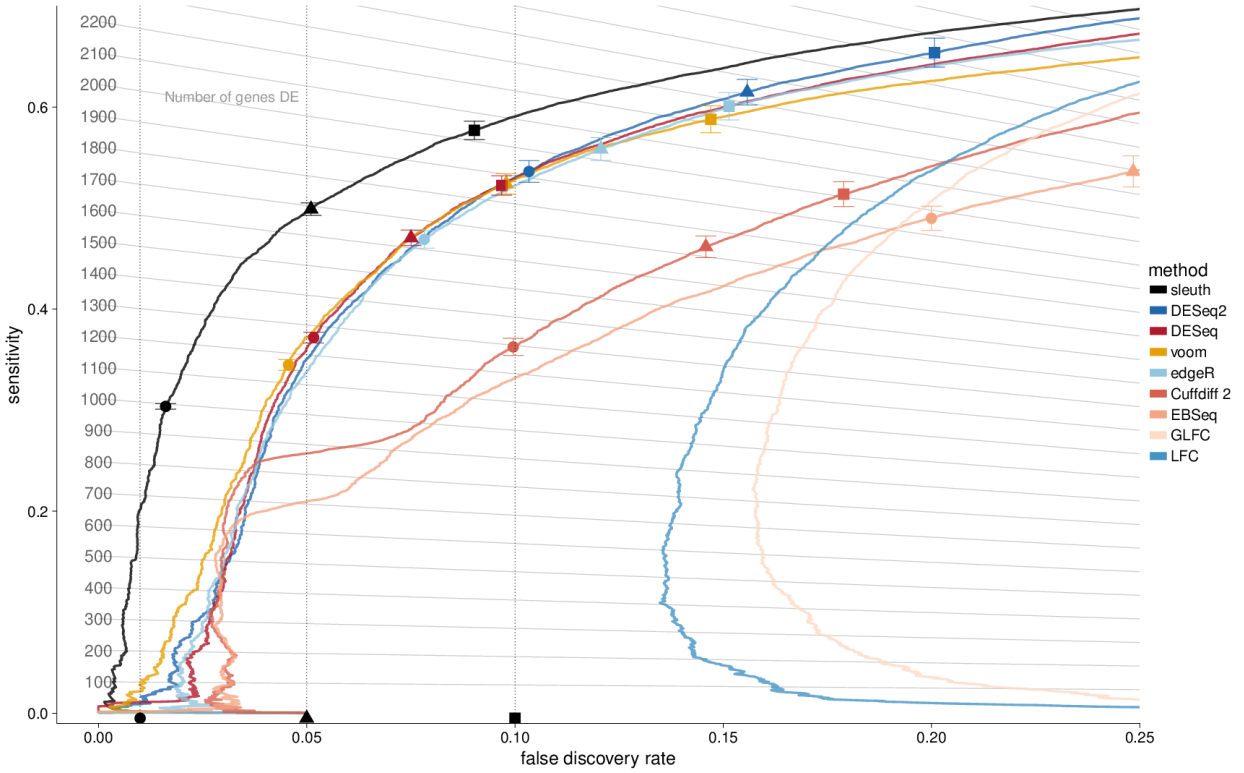
Nature Methods: doi:10.1038/nmeth.4324

4.2.2 Gene level



Supplementary Figure SN9: Zoomed out version of performance on correlated effect simulation at the gene level.

Nature Methods: doi:10.1038/nmeth.4324

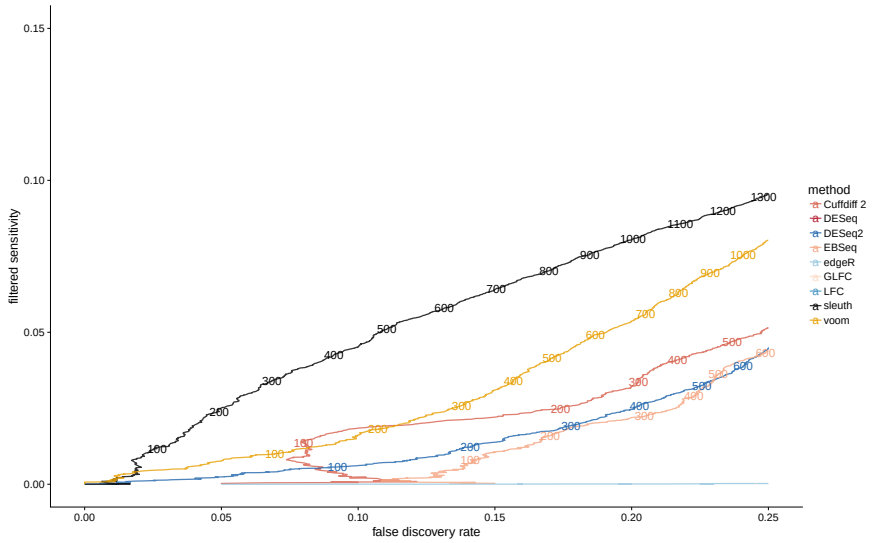


Supplementary Figure SN10: Zoomed in version of performance on correlated effect simulation at the gene level.

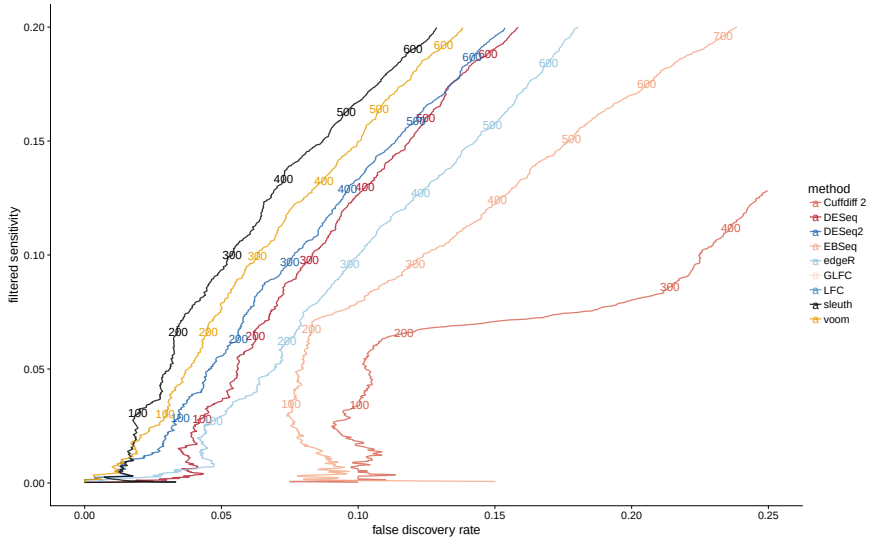
4.2.3 Sensitivity

Our effect-from-experiment simulation includes effect sizes which are small relative to the log fold changes which are typically considered biologically relevant. Because these true positives will be hard to detect, this will result in lower sensitivity. To test performance with larger effect sizes, we show here the result of performing the same analysis as in Figure 2 but where sensitivity is calculated only relative to true positives where the log fold change is at least 2. The results of this are displayed in Supplementary Figure SN11

Nature Methods: doi:10.1038/nmeth.4324



(a) Isoform level sensitivity.



(b) Gene level sensitivity.

Supplementary Figure SN11: Filtering the sensitivity by effect size in the effect from reference simulation. In both figures, the fold change is required to be at least 2 to be considered "positive." The numbers on each line represent the average number of features called differentially expressed (similar to isolines in the other FDR-sensitivity plots.)

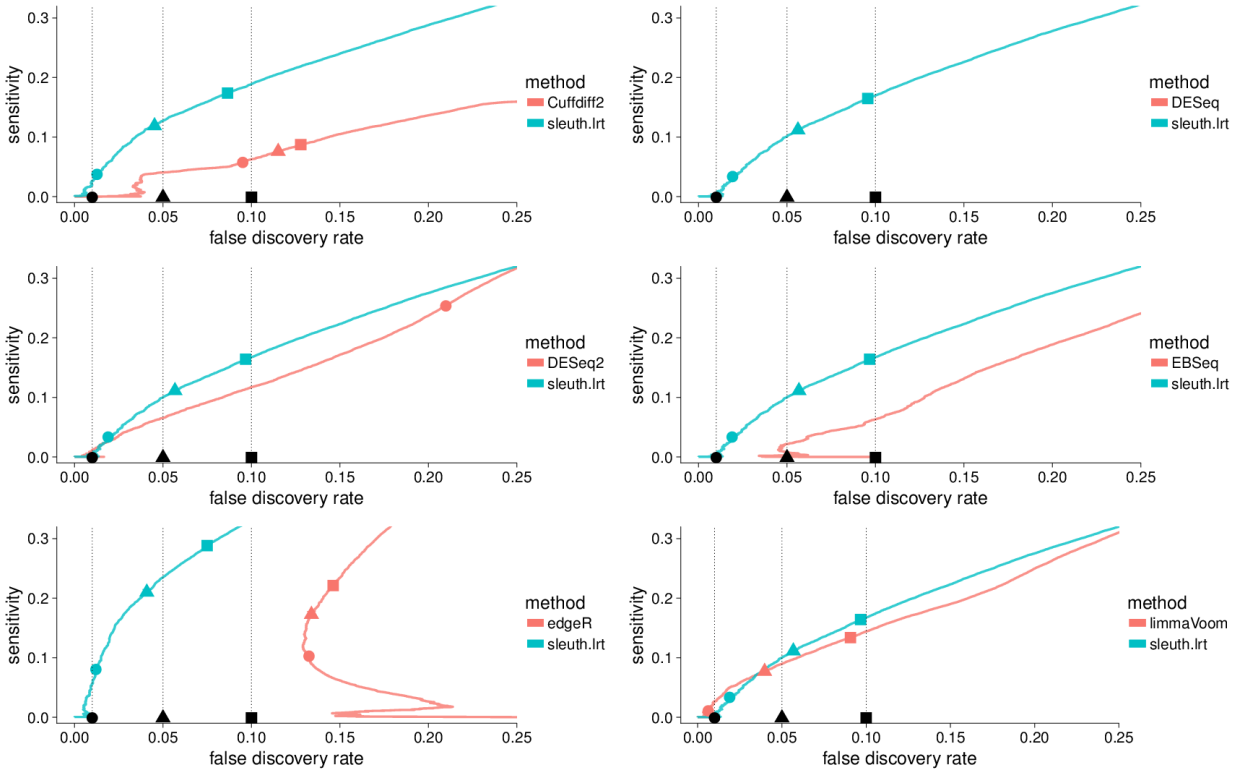
Nature Methods: doi:10.1038/nmeth.4324

5 Performance with common filtering

This section shows the comparison of sleuth to other methods when each pair (sleuth and another method) are tested with a common filter based on intersecting the filtering criteria of both programs. In each case, the two methods were trained using only the data passing both filters. This results in higher power for both methods as there are less tests and fewer low count, high variance targets disrupting the shrinkage estimation.

5.1 Independent effect simulation

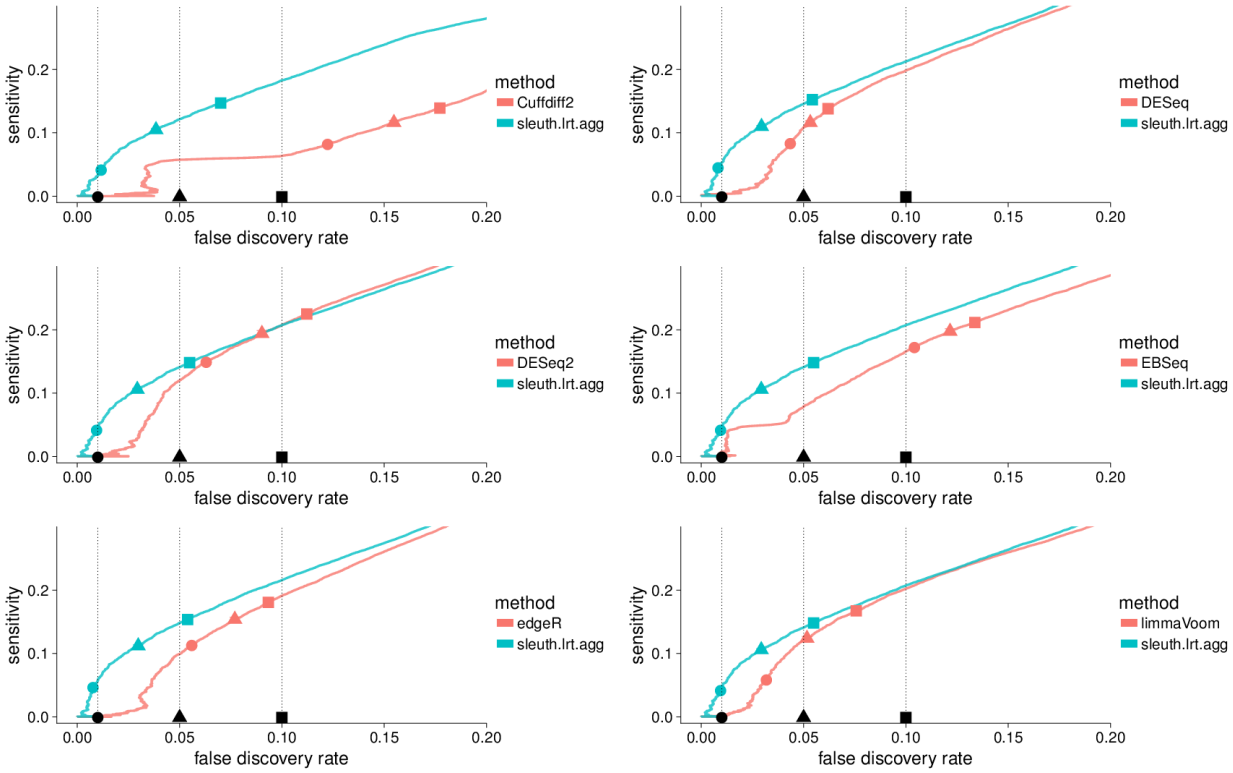
5.1.1 Isoform level



Supplementary Figure SN12: Pairwise comparisons in the independent effect simulation at isoform level with common filtering (DESeq did not register a datapoint in the FDR-sensivity range).

Nature Methods: doi:10.1038/nmeth.4324

5.1.2 Gene level

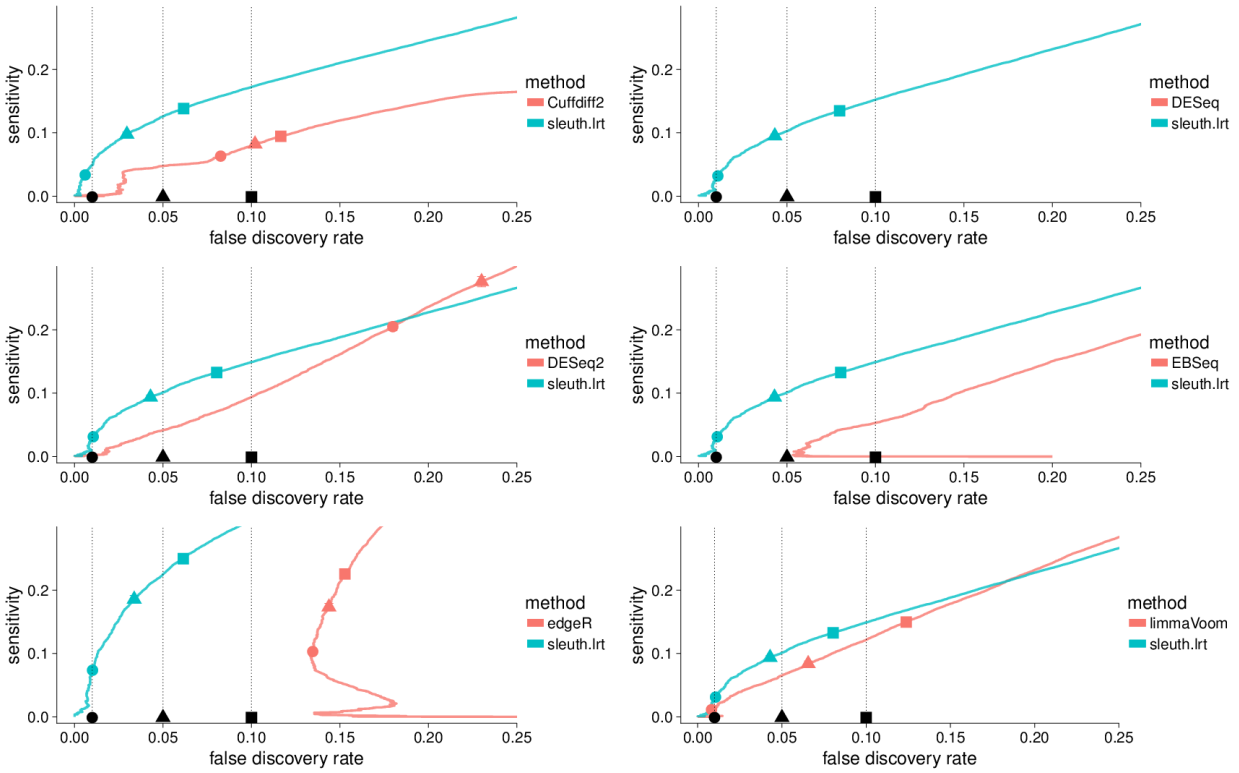


Supplementary Figure SN13: Pairwise comparisons in the independent effect simulation at gene level with common filtering.

Nature Methods: doi:10.1038/nmeth.4324

5.2 Correlated effect simulation

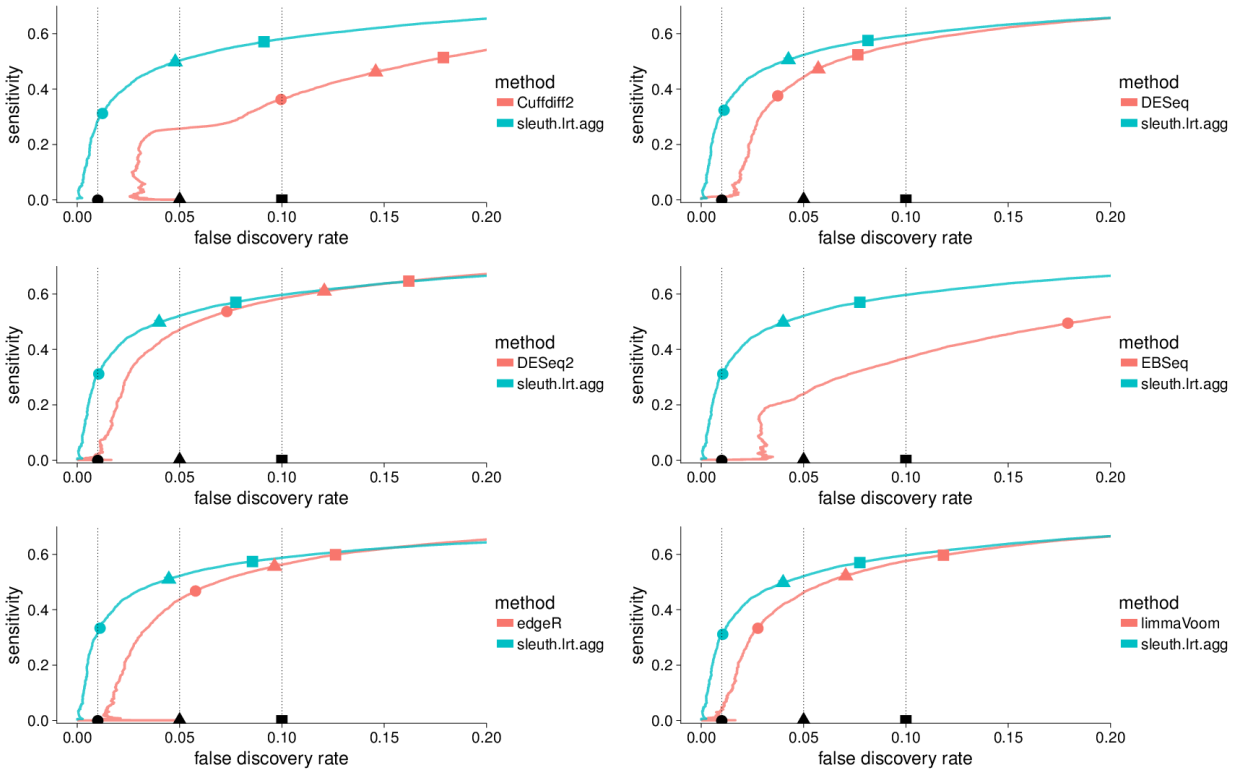
5.2.1 Isoform level



Supplementary Figure SN14: Pairwise comparisons in the correlated effect simulation at isoform level with common filtering (DESeq did not register a datapoint in the FDR-sensivity range).

Nature Methods: doi:10.1038/nmeth.4324

5.2.2 Gene level

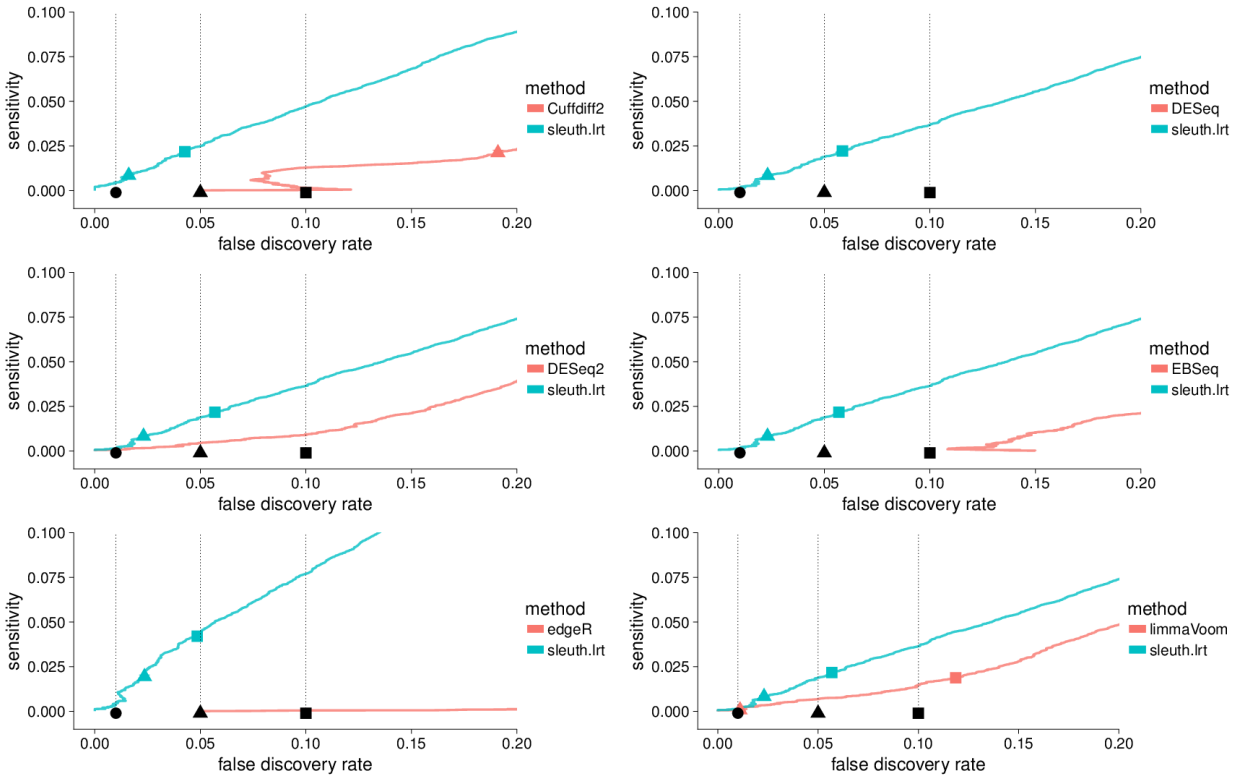


Supplementary Figure SN15: Pairwise comparisons in the correlated effect simulation at gene level with common filtering.

Nature Methods: doi:10.1038/nmeth.4324

5.3 Effect from experiment simulation

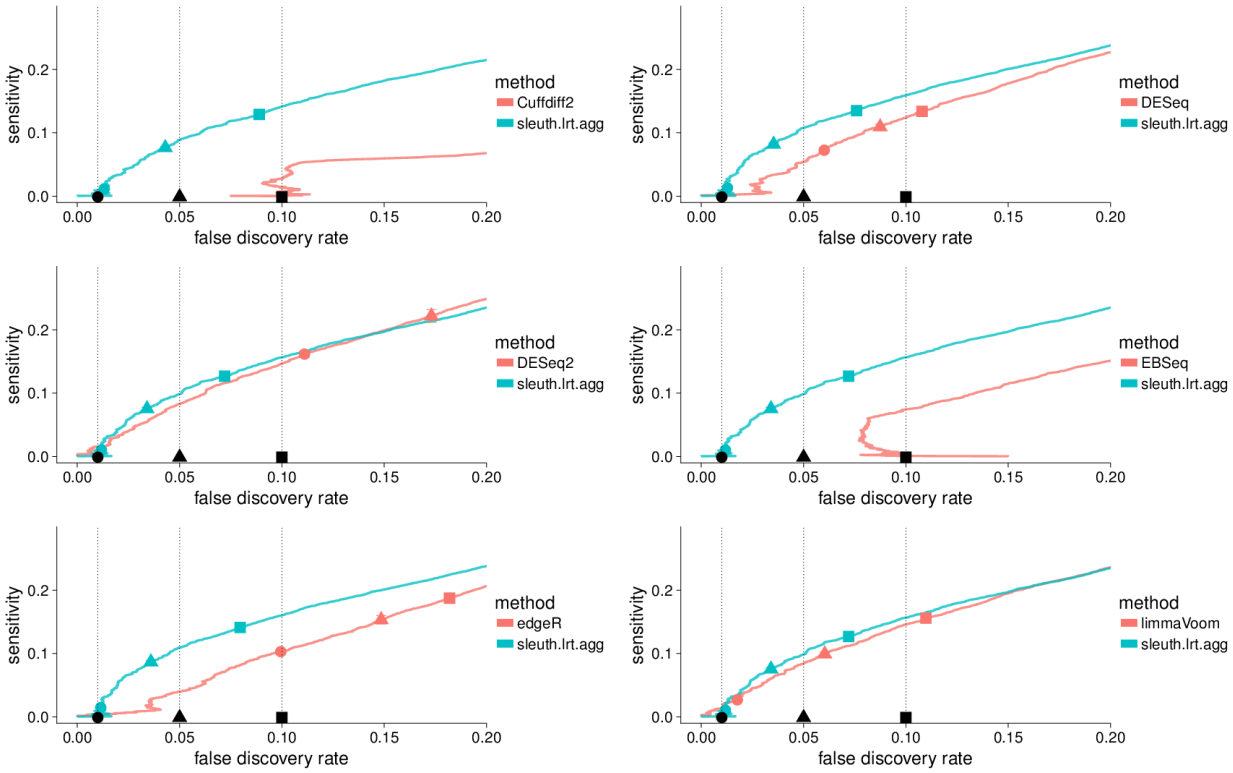
5.3.1 Isoform level



Supplementary Figure SN16: Pairwise comparisons in the effect from experiment simulation at isoform level with common filtering (DESeq did not register a datapoint in the FDR-sensivity range).

Nature Methods: doi:10.1038/nmeth.4324

5.3.2 Gene level

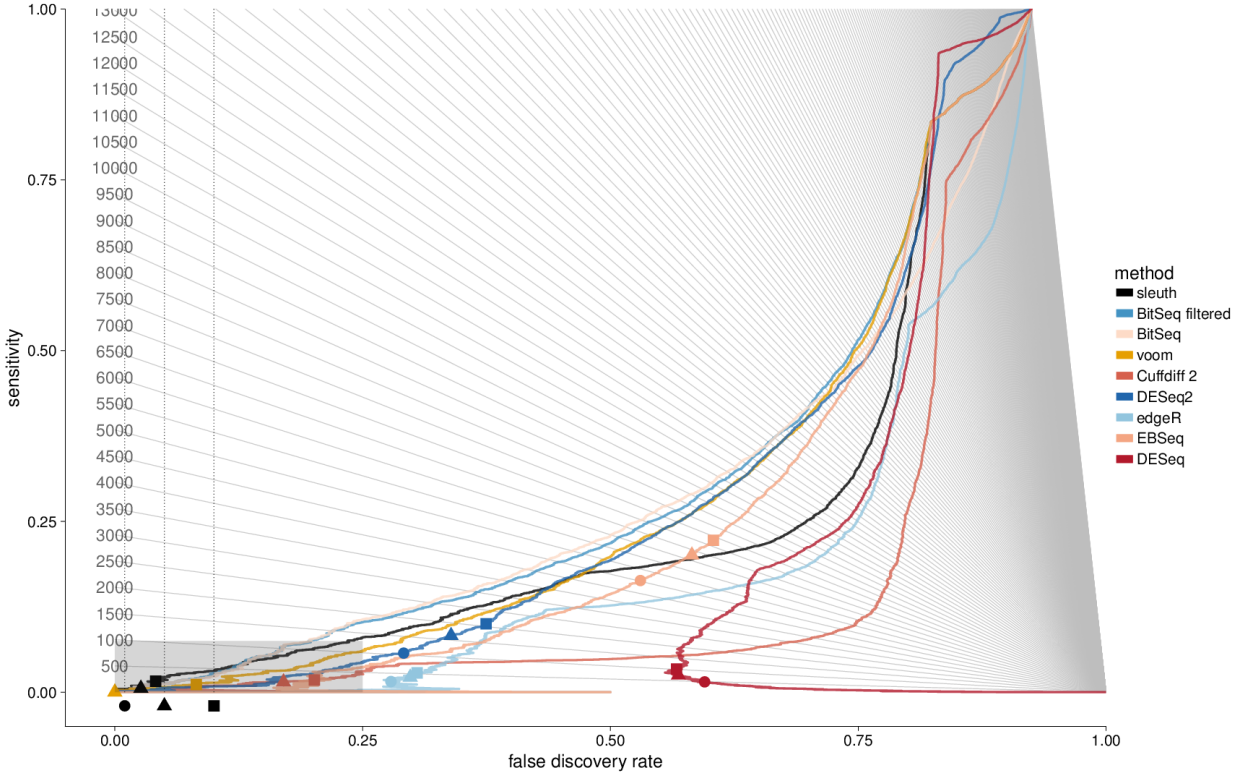


Supplementary Figure SN17: Pairwise comparisons in the effect from experiment simulation at gene level with common filtering.

Nature Methods: doi:10.1038/nmeth.4324

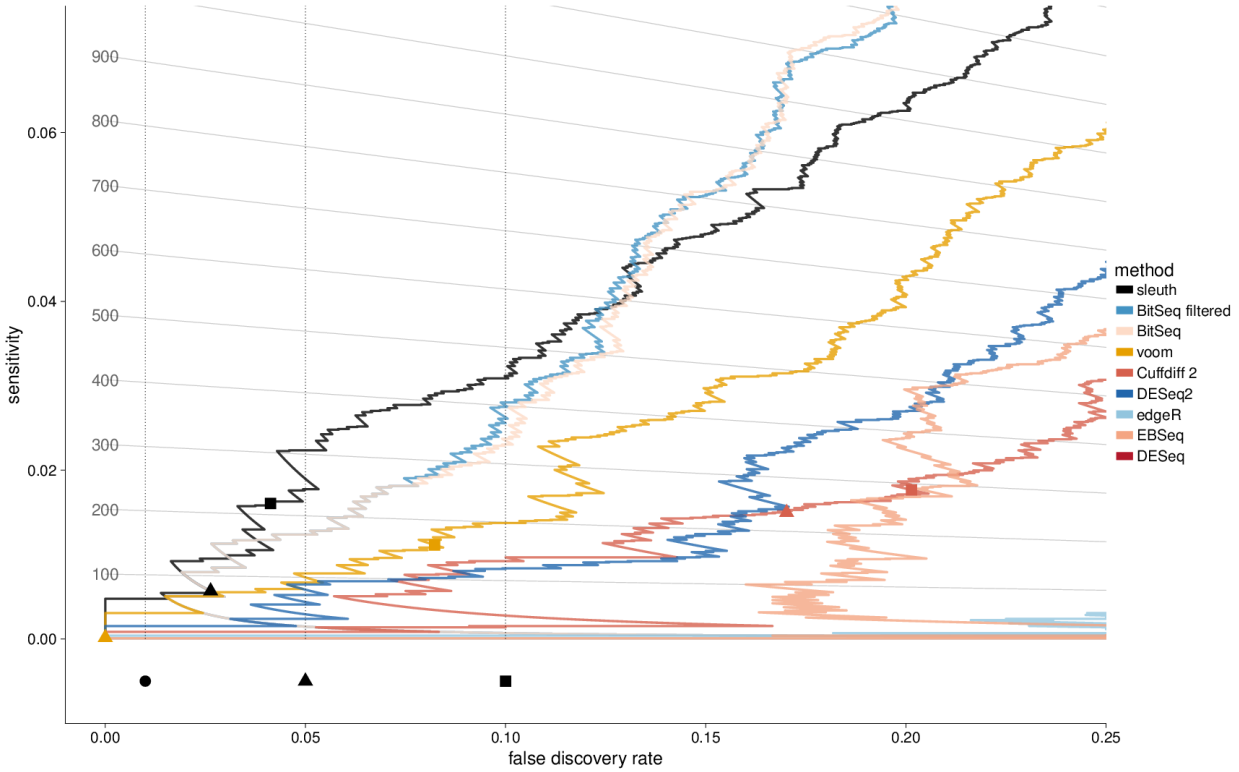
6 Effect from experiment - BitSeq

We only ran BitSeq on one sample due to the long run time. BitSeq does not allow a external filtering method, but we attempted to increase power by intersecting the results with the sleuth filter (BitSeq filtered).



Supplementary Figure SN18: Zoomed out version of performance on independent effect simulation number 1 at the isoform level including BitSeq. Each method was run with the filter provided in their respective manual.

Nature Methods: doi:10.1038/nmeth.4324

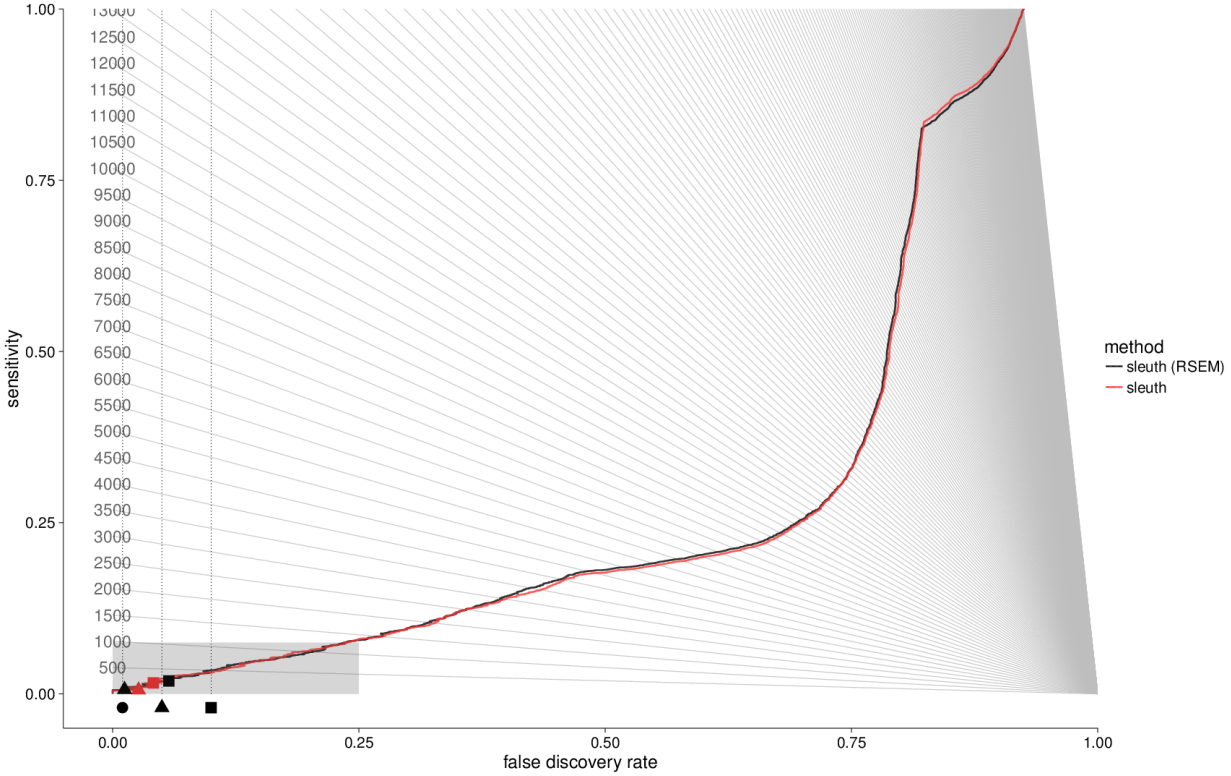


Supplementary Figure SN19: Zoomed in version of performance on independent effect simulation number 1 at the isoform level including BitSeq. Each method was run with the filter provided in their respective manual.

Nature Methods: doi:10.1038/nmeth.4324

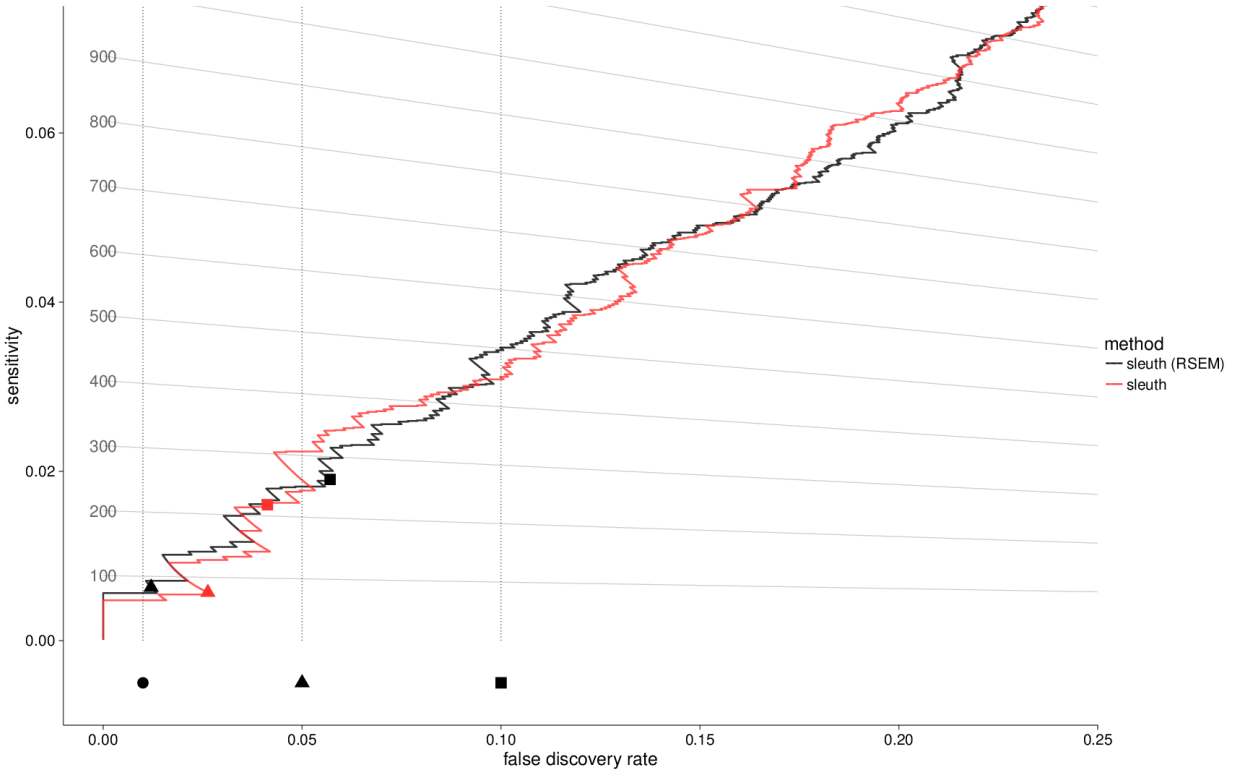
7 Robustness of bootstraps

To check whether or not sleuth is performing better due to kallisto or the new method, we computed bootstraps with RSEM by resampling the data 30 times on one sample. We then read the bootstraps into sleuth and compared how kallisto-sleuth and RSEM-sleuth performed. Supplementary Figures SN20 and SN21 show that the behavior of sleuth with kallisto and RSEM is very similar.



Supplementary Figure SN20: Zoomed out version of performance on independent effect simulation number 1 at the isoform level looking at kallisto-sleuth and RSEM-sleuth.

Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN21: Zoomed in version of performance on independent effect simulation number 1 at the isoform level looking at kallisto-sleuth and RSEM-sleuth.

8 Null experiment

We performed a null experiment to see the number of false positives reported in a real data set. We constructed the null set by using Finnish females from the GEUVADIS data set from sequencing centers that had at least 6 samples [2]. Note that LFC and GLFC were excluded from this analysis as they do not provide false-discovery rate estimates. The experiment is as follows:

- Select a sequencing center.
- Randomly select 3 samples to have a label 'A' and randomly select 3 other samples to have label 'B', all from the same sequencing center chosen in the previous step.
- Call differential expression.

Supplementary Figures 3 and 4 show the number of false positives at the isoform and gene level, respectively.

9 Other variance estimation methods

To demonstrate why we chose taking the max of the smooth estimate and the raw estimate, as well as justifying our model, we include the performance of:

Nature Methods: doi:10.1038/nmeth.4324

- introducing Poisson inferential variance.
- introducing zero inferential variance.
- always taking the smooth estimate rather than the max.

We evaluated these methods on the effect from reference simulation and found that, our method outperforms the other methods. The results can be seen in Supplementary Figures 5 and 6.

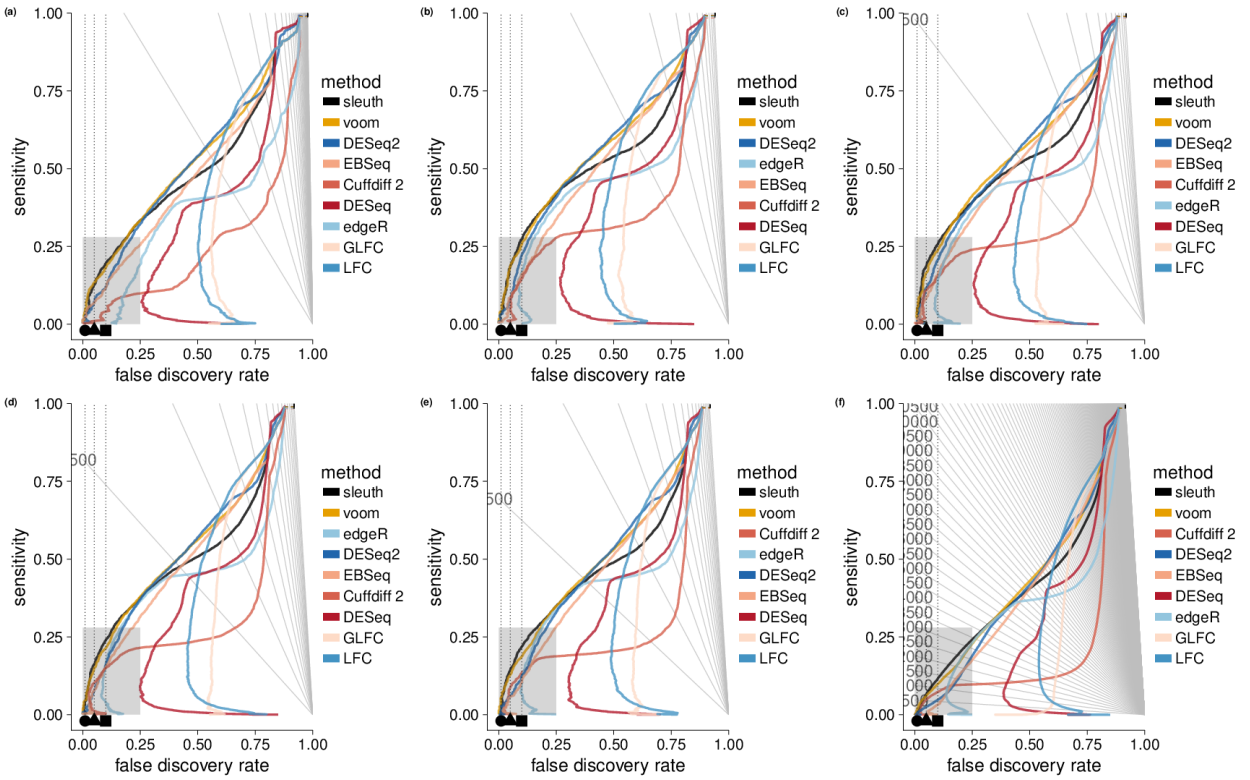
10 Comparison to tximport

We also substituted tximport with kallisto quantification for featureCounts at the gene level for DESeq2, edgeR, and voom to see the effect of changing the gene quantification mode. The results of this experiment are shown in Supplementary Figures 7 and 8. We saw a modest improvement in all of the methods, but sleuth still outperforms other methods on average in the FDR ranges that are most common used.

11 Stratified results

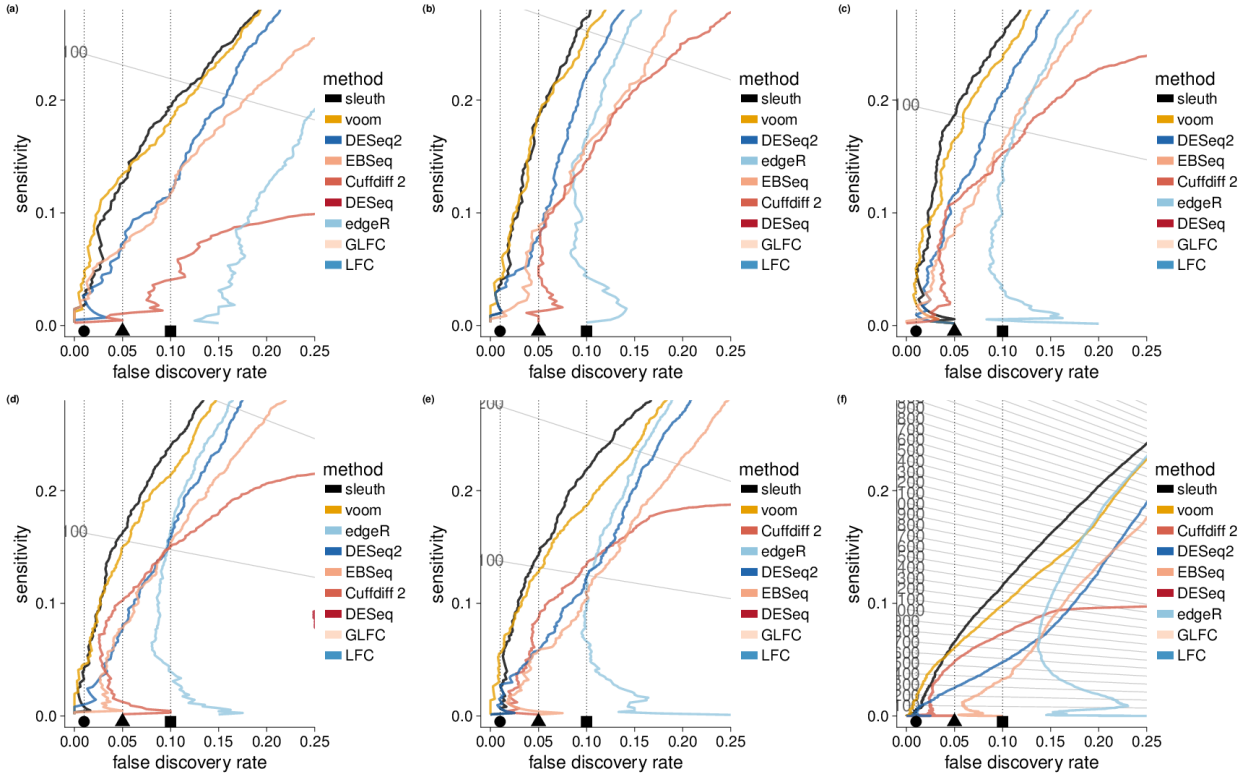
To evaluate how the methods perform with differing gene complexity, we stratified the isoform level simulation by number of isoforms per gene. We see that sleuth performance increases relative to other tools as isoform complexity increases (Supplementary Figures SN22 and SN23).

Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN22: Zoomed out version of performance on independent isoform simulation at the isoform level stratified by number of isoforms per gene. (a), (b), ..., (f) contain isoforms with 1, 2, ..., 6+ isoforms per gene.

Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN23: Zoomed in version of performance on independent isoform simulation at the isoform level stratified by number of isoforms per gene. (a), (b), ..., (f) contain isoforms with 1, 2, ..., 6+ isoforms per gene.

Nature Methods: doi:10.1038/nmeth.4324

	A1	A2	A3	B1	B2	B3
1	25804803	25470509	26167030	29341697	29519936	29488175
2	25885277	25249979	25389240	32370503	30343889	30738282
3	26120446	25827345	25899623	30713524	30243242	30751697
4	25358596	25751655	25122351	33060914	32410330	32989822
5	25256842	25311042	25593932	31890641	30658059	30919290
6	25889040	25051039	26038569	29281347	28643895	28669578
7	25999345	25722280	25292680	30984262	30684305	30958786
8	25601571	25532946	25772470	30649733	30345697	30902790
9	25753144	25161751	26544777	31186168	29804316	30045082
10	26108266	25569269	25855553	32841294	32708783	33222170
11	25742811	25627297	25596448	29374021	28834454	28869359
12	25364068	25137212	25946183	29745829	29789553	30543497
13	25634414	25444127	25682130	27982129	29518922	30585275
14	25090260	25645622	25095267	30182778	29152706	29049277
15	25835121	25671692	25224938	30470838	30650333	30239396
16	25852931	26191755	25509063	31266671	29757417	29669651
17	25615156	25575068	25406585	29418611	30004286	29928450
18	25524970	25356389	26081851	31500275	30436352	30703447
19	26546122	25225294	25908252	30499047	30497149	31153162
20	25949790	25832498	26399110	30414105	29711824	29007750

Supplementary Table 2: Total number of reads for each sample in the independent effect simulation.

Nature Methods: doi:10.1038/nmeth.4324

	A1	A2	A3	B1	B2	B3
1	75312822	25593373	25235809	9839136	9929876	90460751
2	76754874	77164818	8448302	10128818	91488796	10260035
3	25605293	8695919	25491639	88543891	91209909	10263281
4	78615038	8506970	8385197	9938514	92448519	89436730
5	77364271	74865995	8528707	10076750	10042850	88754945
6	25860713	76275460	8440966	9610553	89957747	29271608
7	8597415	25281982	75251435	10213462	30323374	95540321
8	25522744	76270625	25088664	10137245	10421655	91773477
9	8666259	76943845	8559758	9987635	88007539	89331914
10	8566416	25410620	8490243	30555362	90843296	91010457
11	8455041	75784558	25372293	30267698	87570393	9787472
12	8492502	78575682	77077410	29654868	29818839	9886687
13	25457086	26158929	8471727	29118444	29471345	90230274
14	26002373	26260610	25956110	97485646	31266203	10420089
15	25487509	8493425	25243386	87415490	88713112	9713101
16	77486094	8448668	8663999	29099399	86839687	28969442
17	8407592	25888803	76954183	10420214	30992646	94468557
18	25873615	76371611	8359367	31089875	10122162	88988075
19	25678980	8501752	76183257	10537835	91483251	32587127
20	8454584	25950638	78035981	10428599	30550328	91394732

Supplementary Table 3: Total number of reads for each sample in the correlated effect simulation.

Nature Methods: doi:10.1038/nmeth.4324

	A1	A2	A3	B1	B2	B3
1	75168780	25485734	8384121	25747475	8534405	77027331
2	8501164	76807800	77633767	8740185	8684330	79993067
3	25331289	76540098	78511757	8447971	8427418	25987110
4	8513095	77563156	76171563	8746602	8828539	78111712
5	25944623	25748634	78161328	8493836	77658353	8625305
6	74728858	25432069	78836864	8805673	8985756	26342445
7	25774678	76049917	25580674	8843625	9010990	81891037
8	8443669	25980365	76146962	26349727	78607969	8721963
9	8526924	25514760	25393205	8893176	81080795	79541529
10	76685128	25597960	8447489	25772831	8863348	79552345
11	26124390	76923452	8679035	26027252	26363462	26557951
12	77306728	8528461	8413117	8841991	80696745	78431231
13	8467790	8686289	75067527	78678591	8653662	78027739
14	25726894	76768529	8477740	79007076	27257417	8990530
15	8317198	26005481	76363404	8420821	76489073	26140406
16	77110090	76973967	8803533	8982341	8838866	78798474
17	8595280	8660676	78462128	9011452	79474418	78080390
18	78598555	8504087	76359175	8780854	80048666	8739335
19	76495399	25188780	8438441	25570493	25599853	25619140
20	74677076	8417349	25938160	8933135	26341988	79293035

Supplementary Table 4: Total number of reads for each sample in the effect from experiment simulation.

A1	76567345	B1	77353279
A2	25501413	B2	8851265
A3	25201440	B3	26670650
A4	8629446	B4	77175275
A5	77679146	B5	8502371
A6	76799156	B6	26087240
A7	8496819	B7	25635027
A8	77535590	B8	8644561
A9	8575806	B9	26012768
A10	25820678	B10	79047338
		B11	8562183

Supplementary Table 5: Total number of reads for each sample in the effect from experiment simulation used in the self-referential FDR experiment.

Nature Methods: doi:10.1038/nmeth.4324

References

[1] Zhen-Xia Chen, Kseniya Golovnina, Hina Sultana, Satish Kumar, and Brian Oliver. Transcriptional effects of gene dose reduction. *Biology of Sex Differences*, 5(1):5, 2014.

[2] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedlnder, Peter A. C. t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, ngel Carracedo, Stylianos E. Antonarakis, Robert Hsler, Ann-Christine Syvnen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guig, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013.

[3] Mingxiang Teng, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, Christopher E. Mason, Sara Olson, Dmitri Pervouchine, Cricket A. Sloan, Xintao Wei, Lijun Zhan, and Rafael A. Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17:74, 2016.

[4] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2013.

Nature Methods: doi:10.1038/nmeth.4324

Supplementary Note 2: Details of the sleuth model

Harold Pimentel Nicolas Bray Suzette Puente Páll Melsted
Lior Pachter

April 27, 2017

1 The model

We use the term *experiment* to denote the measurement of transcript abundances from a series of n samples which are related by an $n \times p$ design matrix x. Each row vector x_i $(i = 1, \ldots, n)$ of the matrix x records the fixed design characteristics of sample i with respect to the p covariates.

For each transcript t and sample i, we model the logarithm of transcript abundance (measured in counts) with a latent random variable Y_{ti}. A vector β_t of length p associates fixed effects to each transcript, and "biological noise" ϵ_{ti} perturbs the response $x_i^T \beta_t$ so that

$$Y_{ti} \mid x_i = x_i^T \beta_t + \epsilon_{ti}. \tag{1}$$

While many RNA-Seq models posit that Y_{ti} is observed, the ambiguity of read (pseudo)-alignments means that instead what is measured is

$$D_{ti} \mid y_{ti} = y_{ti} + \xi_{ti}, \tag{2}$$

where ξ_{ti} is "inferential noise".

Assuming that ϵ_{ti} and ξ_{ti} are random variables satisfying $\epsilon_{ti} \sim \mathcal{N}(0, \sigma_t^2), \xi_{ti} \sim \mathcal{N}(0, \tau_t^2)$, $\mathrm{cov}(\epsilon_{ti}, \epsilon_{tj}) = \mathrm{cov}(\xi_{ti}, \xi_{tj}) = 0 \ \forall i \neq j, \mathrm{cov}(\epsilon_{ti}, \xi_{tj}) = 0 \ \forall i, j$ and that $\forall t \neq u$, ϵ_t, ξ_t are independent of ϵ_u and ξ_u respectively, we have that $Y_t = (Y_{t1}, Y_{t2}, ..., Y_{tn})$ and $D_t = (D_{t1}, D_{t2}, ..., D_{tn})$ are both normally distributed as

$$Y_t \sim \mathcal{N}(x\beta_t, \sigma_t^2 I_n), \tag{3}$$
$$D_t \sim \mathcal{N}(x\beta_t, (\sigma_t^2 + \tau_t^2) I_n). \tag{4}$$

This model is known as the *response error measurement model* with no error on the covariates [3]. For completeness, we describe some of its properties below and explain how they apply to parameter estimation in the context of the sleuth workflow.

2 Overview of sleuth workflow

The input to sleuth consists of estimated counts for transcripts in the samples constituting the experiment as well as estimates of variance for those counts obtained from bootstraps.

Nature Methods: doi:10.1038/nmeth.4324

Both the estimated counts and the variance are output by kallisto. The (estimated) counts for transcript t in sample i is referred to as c_{ti} and the variance of D_{ti} given y_{ti} estimated from the boostraps of kallisto. The sleuth workflow begins with a filtering of low abundance transcripts, followed by the application of two normalizations and then parameter estimation for the model described above. This enables the regularization of the biological variance contributing to transcript abundance variance across samples, and finally to an overall total variance estimate for each transcript. The workflow can be applied to either transcripts, or groups of transcripts such as genes, and the two modes are described below.

3 Filtering prior to parameter estimation

Prior to estimating parameters of the model we filter low abundance transcripts. This helps in fitting the model. We ignore transcripts where there are less than 5 estimated counts in more than 47% of the samples, i.e. when $|\{i : c_{ti} < 5\}| \geq 0.47 \cdot n$.

4 Normalization and transformation

Following the filtering there are two different normalizations that we apply to the estimated counts c_{ti}: between sample normalization and within sample normalization. First, we perform between sample normalization to estimate sample specific size factors s_i on the estimated counts following the DESeq procedure [1] applied to transcripts:

$$\hat{s}_i = \text{median}_t \frac{c_{ti}}{\left(\prod_{j=1}^{n} c_{tj}\right)^{\frac{1}{n}}}.$$

Following between sample normalization, we log transform the data so that transcripts have similar variance across samples.

For each transcript, the abundance is estimated as the (normalized) log estimated count

$$\begin{aligned} d_{ti} &= \log\left(\frac{1}{\hat{s}_i}\tilde{l}_{ti}\frac{c_{ti}}{\tilde{l}_{ti}} + 0.5\right) \\ &= \log\left(\frac{1}{\hat{s}_i}c_{ti} + 0.5\right), \end{aligned}$$

where $\tilde{l}_{ti}$ is the effective length of transcript t in sample i. Note that the expression $\frac{c_{ti}}{\tilde{l}_{ti}}$is proportional to the abundance of transcript t in sample i, and that the multiplication by the effective length serves to rescale the abundance estimate to a count estimate. The offset of 0.5 is used to ensure that the argument to the logarithm is positive.

5 Estimation of β_t

Conveniently, the standard ordinary least squares estimators for the fixed effects are unbiased under this model. The standard estimator is

$$\hat{\beta}_t = (x^T x)^{-1} x^T d_t \tag{5}$$

where $d_t = (d_{t1}, \ldots, d_{tn})$. The expected value of $\hat{\beta}_t$ is

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_t] &= \mathbb{E}[\mathbb{E}[(x^T x)^{-1} x^T d_t \mid y_t]] \\
&= \mathbb{E}[(x^T x)^{-1} x^T \mathbb{E}[y_t + \xi_t | y_t]] \\
&= \mathbb{E}[(x^T x)^{-1} x^T (y_t + 0)] \\
&= \mathbb{E}[(x^T x)^{-1} x^T (x\beta_t + \epsilon_t)] \\
&= (x^T x)^{-1} x^T \mathbb{E}[(x\beta_t + \epsilon_t)] \\
&= (x^T x)^{-1} x^T (x\beta_t + 0) \\
&= \beta_t.
\end{aligned}$$

Thus $\hat{\beta}_t$ is an unbiased estimator of β_t.

6 Estimation of the variance of D_{ti}

The variance of D_{ti} decomposes according to the law of total variance:

$$\begin{aligned}
\mathbb{V}[D_{ti}] &= \mathbb{E}[\mathbb{V}[D_{ti} \mid y_{ti}]] + \mathbb{V}[\mathbb{E}[D_{ti} \mid y_{ti}]] \\
&= \mathbb{E}[\mathbb{V}[y_{ti} + \xi_{ti} \mid y_{ti}]] + \mathbb{V}[y_{ti}] \\
&= \mathbb{E}[\tau_t^2] + \sigma_t^2 \\
&= \tau_t^2 + \sigma_t^2.
\end{aligned}$$

The inferential variance τ_t^2 is estimated as the mean of the sample variance estimates $\hat{\tau}_{ti}^2$ which are obtained from kallisto with the bootstrap:

$$\hat{\tau}_t^2 = \frac{1}{n} \sum_i \hat{\tau}_{ti}^2. \tag{6}$$

Using the second moment as an estimator for the total variance, namely

$$\hat{\mathbb{V}}[D_{ti}] = \frac{1}{n-p} \sum_{i=1}^{n} (d_{ti} - x_i^T \hat{\beta}_t)^2$$

and solving for the (raw) biological variance, we obtain

$$\hat{\sigma}_t^2 = \max\left(\left(\frac{1}{n-p} \sum_{i=1}^{n} (d_{ti} - x_i^T \hat{\beta}_t)^2 \right) - \hat{\tau}_t^2, 0 \right), \tag{7}$$

Nature Methods: doi:10.1038/nmeth.4324

where the max operation is necessary to ensure that the (raw) biological variance is nonnegative.

When $n - p$ is small, the (raw) biological variance estimate $\hat{\sigma}_t^2$ is unstable. Since this is the situation in almost all RNA-Seq studies, we regularize the biological variance estimate by shrinkage. We split the abundance values into 100 "windows" (ranges) such that each window, w, contains 1% of the mean abundance under the intercept only model. Note that each transcript t has an abundance contained in a single window denoted by $w(t)$ and a distribution of estimated biological variance is associated to that window, namely of $\hat{\sigma}_t^2$. We denote by $IQR(w(t))$ the interquartile range of the distribution associated to a window $w(t)$ and identify a training set of transcripts R for shrinkage using these interquantile ranges:

$$R = \{t : \hat{\sigma}_t^2 \in IQR(w(t))\}. \tag{8}$$

We perform LOESS on the set R and perform shrinkage on the square root of the standard deviation similar to voom [4] as this results in more stable estimates. Our shrunken estimate of σ_t^2 is then a function of the mean $\bar{d}_t = \frac{1}{n}\sum_{i=1}^{n} d_{ti}$ (the parameter estimate under the intercept only model):

$$\tilde{\sigma}_t^2 = f(\bar{d}_t) = \left[(\text{loess}_{r\in R}(\bar{d}_r, \hat{\sigma}_r^{\frac{1}{2}}))(\bar{d}_t)\right]^4. \tag{9}$$

Our final estimate of the total variance of transcript t in sample i is therefore (the sample independent expression)

$$\hat{\mathbb{V}}[D_{ti}] = \max(\tilde{\sigma}_t^2, \hat{\sigma}_t^2) + \hat{\tau}_t^2. \tag{10}$$

7 Gene level estimates

The sleuth model for transcript abundance, and the associated parameter estimation described above can be generalized to groups of transcripts such as genes. To do so, we first note that a set of genes can be viewed as a partition of the set of transcripts, so that each gene g is just a set of transcripts. To model gene abundances, we replace transcript abundance with gene abundance in the model as follows:

Starting with the same design matrix x as in the transcript case, for each gene g and sample i, we model the logarithm of transcript abundance (measured in counts) with a latent random variable Y_{gi}. A vector β_g of length p associates fixed effects to each gene, and "biological noise" ϵ_{gi} perturbs the response $x_i^T\beta_g$ so that

$$Y_{gi} \mid x_i = x_i^T\beta_g + \epsilon_{gi}. \tag{11}$$

As before, we use a response error measurement model based on underlying normality assumption which leads to

$$Y_g \sim \mathcal{N}(x\beta_g, \sigma_g^2 I_n), \tag{12}$$

$$D_g \sim \mathcal{N}(x\beta_g, (\sigma_g^2 + \tau_g^2) I_n). \tag{13}$$

The workflow at the gene level is identical to that of transcript level analysis, with a few minor differences:

(a) At the gene level, if at least one isoform in that gene passes the filter, the entire gene passes the filter.

(b) The normalization at the gene level is analogous to that at the transcript level except for two differences: the abundance of genes is first calculated by summing up the abundances of the constituent isoforms and the effective length of a single transcript is replaced by an effective length for the gene (consisting of the median of the effective lengths of the constituent transcripts). For a gene G the normalized estimate for abundance in "effective counts" is therefore

$$d_{gi} = \log\left(\frac{1}{\hat{s}_i}(\text{median}_{t\in G}\,\tilde{l}_{ti})\sum_{t\in G}\frac{c_{ti}}{\tilde{l}_{ti}} + 0.5\right).$$

(c) The shrinkage procedure is applied at the gene level, leading to a total variance estimate of

$$\hat{\mathbb{V}}[D_{gi}] = \max(\tilde{\sigma}_g^2, \hat{\sigma}_g^2) + \hat{\tau}_g^2,$$

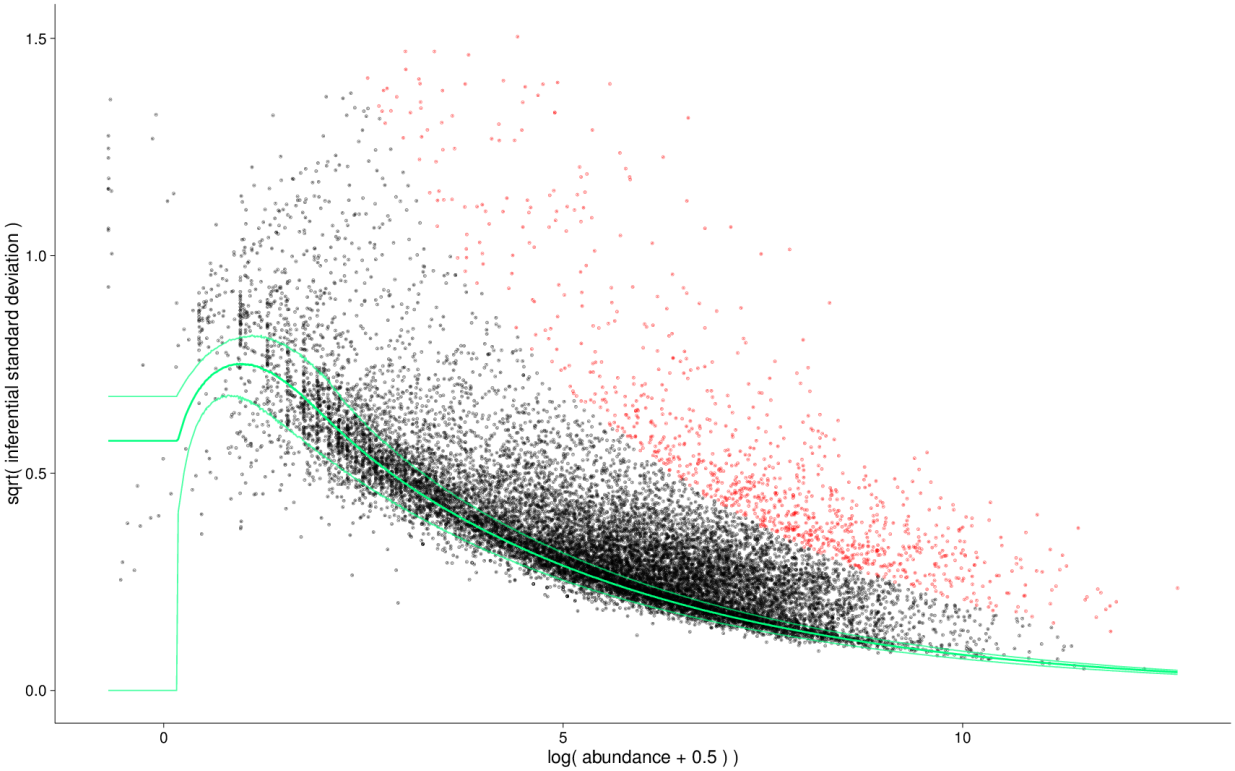
where the estiamtes of $\hat{\sigma}_g^2$ and $\hat{\tau}_g^2$ are analogous to their transcript counterparts.

Note that d_{gi} reduces to d_{ti} when g consists of just a single isoform, so that the workflow can be viewed as a direct generalization of the transcript level case. Moreover, the procedure outlined above can be applied to sets of transcripts obtained from any partition of the transcriptome.

8 Decomposition of the variance

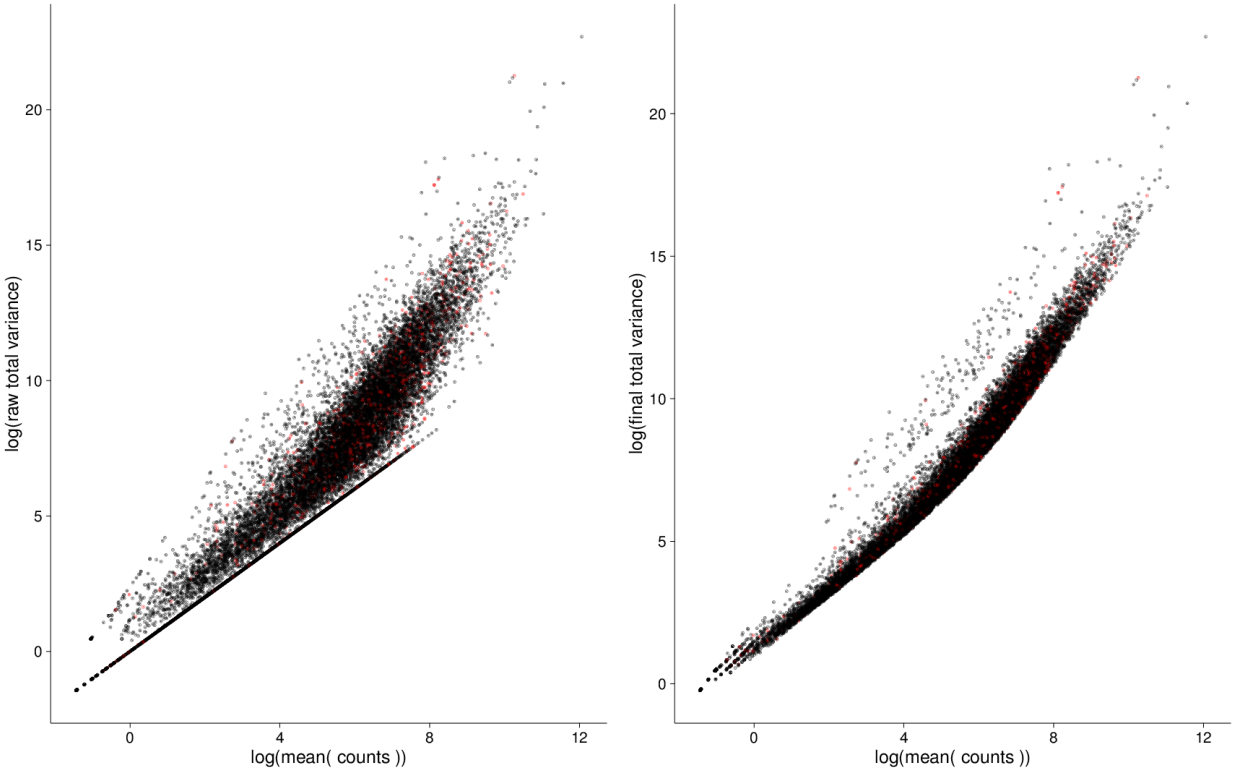
To demonstrate the decomposition of variance as performed by sleuth, we here examine a gene-level analysis of data from Bottomly et al. [2]. Genes with high inferential variance (red dots in Supplementary Figure SN1) are lumped together with genes with high biological variance by DESeq2 (Supplementary Figure SN2). This makes it difficult to correctly assign a high total variance to those genes prior to differential analysis (Supplementary Figure SN3). Unlike DESeq2, by decoupling biological and inferential variance (Supplementary Figures SN3, SN4), sleuth assigns a high total variance to most of the genes with high inferential variance (Supplementary Figure SN5).

Nature Methods: doi:10.1038/nmeth.4324



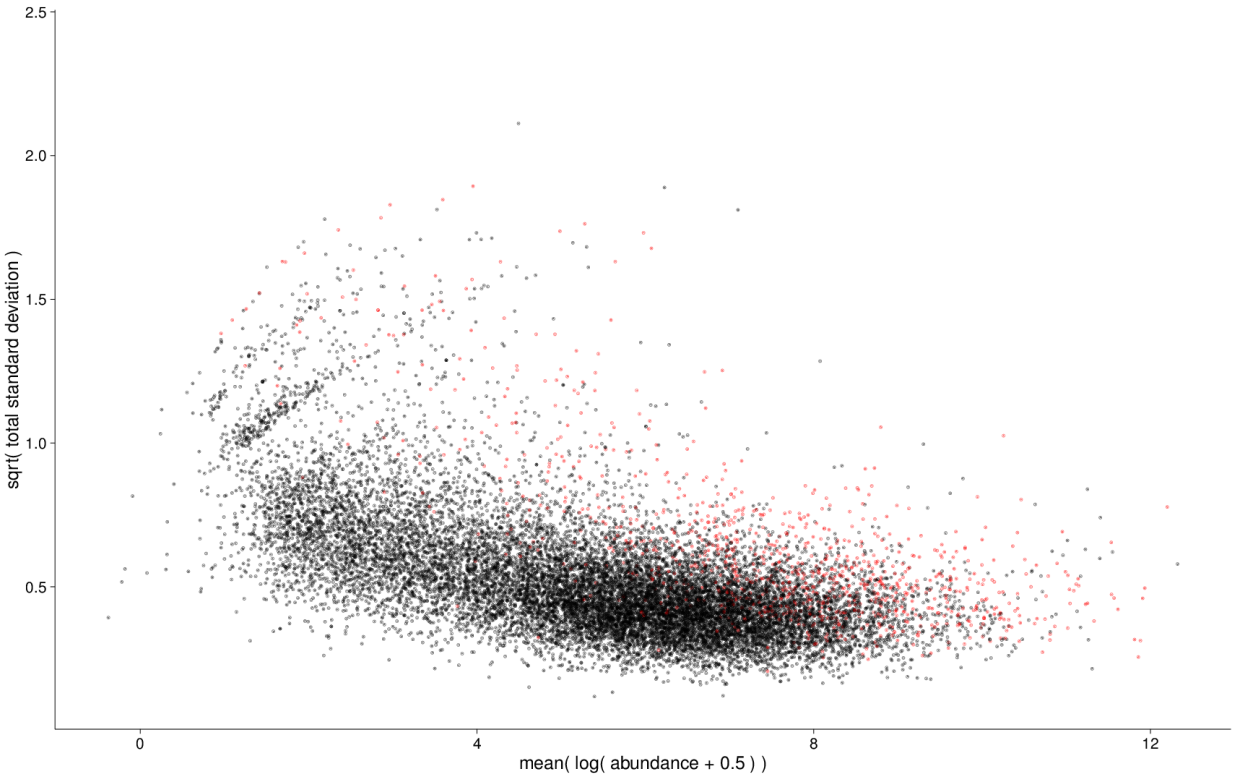
Supplementary Figure SN1: Inferential variance on sample SRR099228. The x-axis is the gene abundance, and the y-axis is the bootstrap estimate of the inferential variance. The green lines represent the 5% confidence bound, mean, and 95% confidence bound expected under the Poisson model.

Nature Methods: doi:10.1038/nmeth.4324



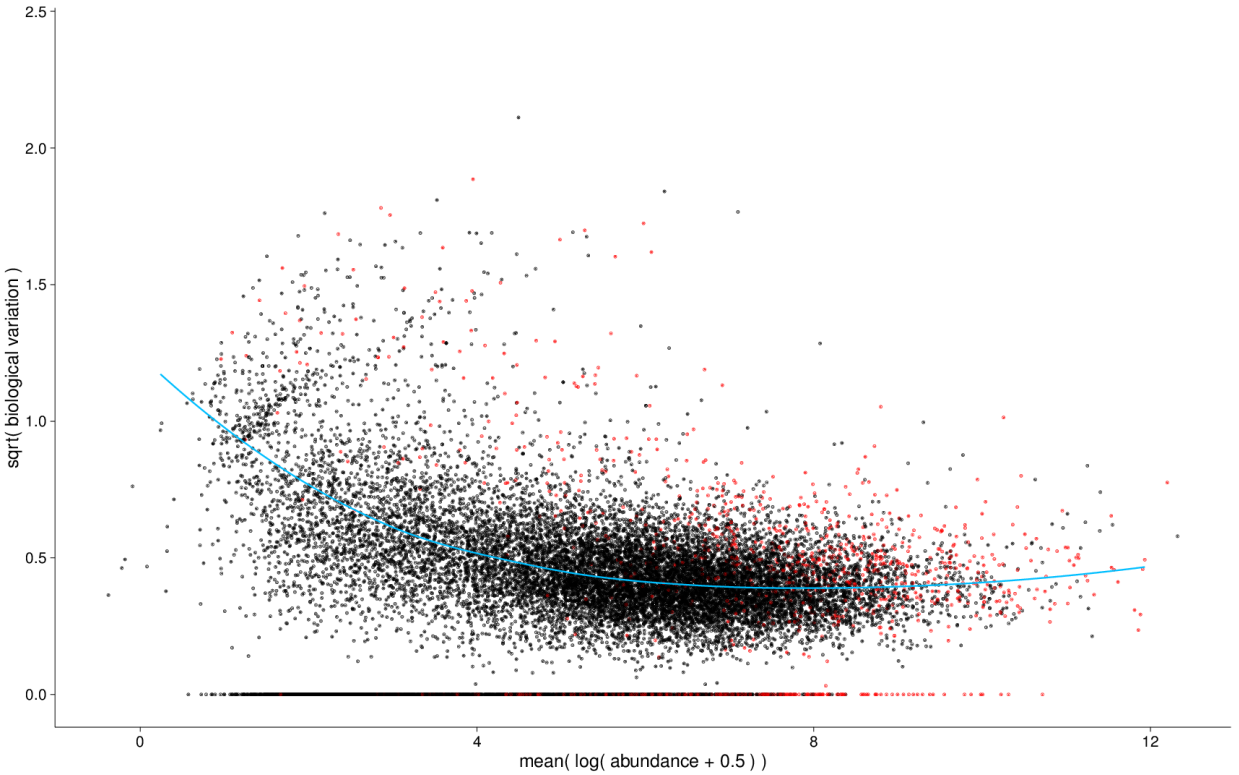
Supplementary Figure SN2: Mean expression versus total variance as estimated by DESeq2. The left panel contains the raw estimates of the variance. The right panel contains the smoothed estimate of the variance. Note that the outliers are fairly randomly distributed across variance and expression patterns.

Nature Methods: doi:10.1038/nmeth.4324



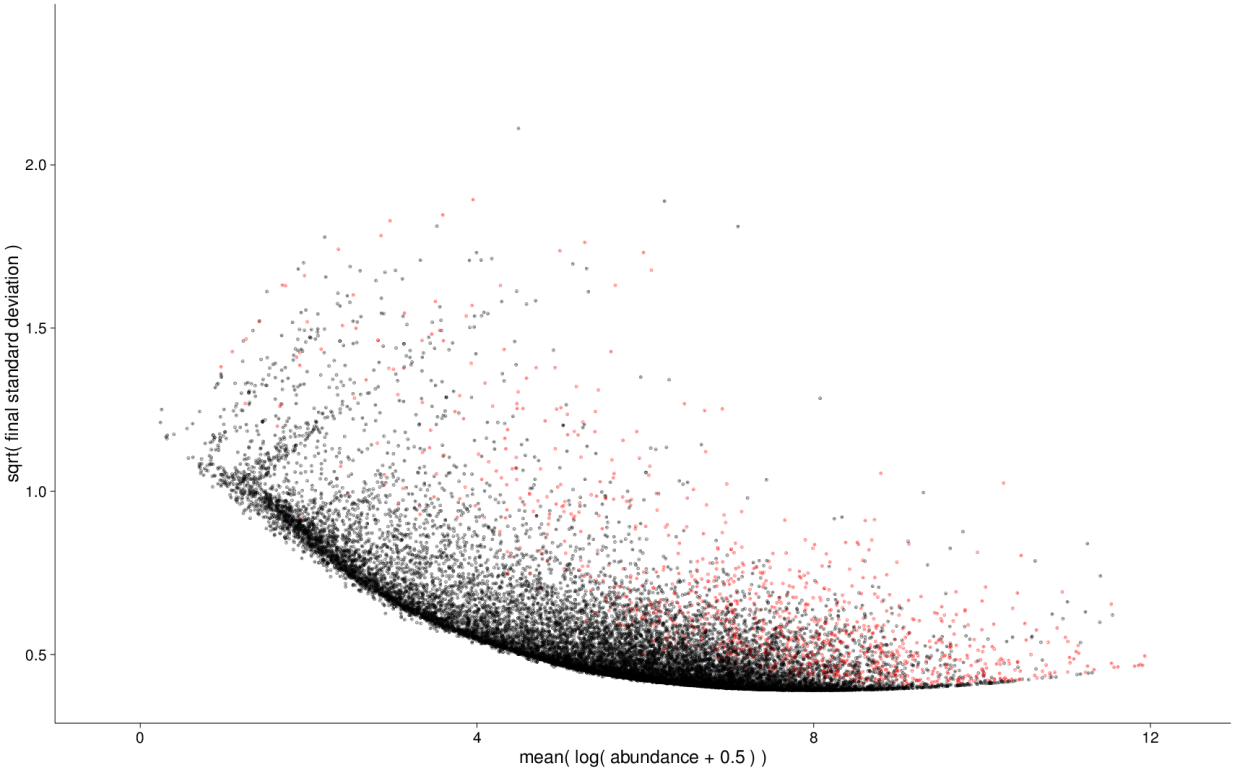
Supplementary Figure SN3: Raw total variance as estimated by sleuth. The outliers are fairly randomly distributed since they do not consider the inferential variance.

Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN4: Biological variance as estimated by sleuth once the inferential variance has been removed. The blue line represents the mean-biological variance relationship modeled in sleuth. Note that in many cases the inferential variance is greater than the biological variance resulting in an estimate of biological variance equal to zero.

Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN5: Final total variance as modeled by sleuth. Note that almost all of the outliers have higher abundance than the non-outliers due to high inferential variance.

References

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[2] Daniel Bottomly, Nicole A. R. Walter, Jessica Ezzell Hunter, Priscila Darakjian, Sunita Kawane, Kari J. Buck, Robert P. Searles, Michael Mooney, Shannon K. McWeeney, and Robert Hitzemann. Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *PLOS ONE*, 6(3):e17820, March 2011.

[3] John P Buonaccorsi. *Measurement error: models, methods, and applications*. CRC Press, 2010.

[4] Charity W Law, Yunshen Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, 2014.

Nature Methods: doi:10.1038/nmeth.4324

Supplementary Note 3: Interactive Exploratory Data Analysis Using sleuth

Harold Pimentel Nicolas Bray Suzette Puente Páll Melsted
Lior Pachter

April 27, 2017

The interpretation and analysis of RNA-Seq data is complicated by a number of factors: the large number of reads sequenced in typical experiments (many millions) and the large number of transcripts/genes under consideration (typically tens of thousands) make it difficult to interactively examine the data. However exploratory data analysis is important both for understanding how to analyze data and in the formulation of hypotheses about the results. To address this issue, and to make it easier to interpret and assess results from sleuth, we have developed a Shiny-based [**?**] interactive app called *sleuth live* for examining sleuth results.

Sleuth live allows for several types of analyses instantly. While other software packages have plotting capabilities [**?**], this often requires reading the documentation then choosing the proper arguments to make each plot. Sleuth provides this functionality for users who want it, but sleuth live automatically makes the majority of the plots that are made in most analyses. Additionally, many of the plots are interactive so that users can zoom in on interesting regions. If a user wants to look at a different facet of the data (e.g. different sample, color by a different variable), sleuth live automatically pre-populates valid arguments in drop-down menus where possible. Changes in the drop-down menus immediately result in updated plots. Finally, most plots and tables can be exported.

Currently, the following analyses are provided automatically in sleuth live:

- heatmap of specified transcripts/genes.
- transcript/gene plots of sample specific quantification with inferential variance.
- MA plots.
- table of summary statistics per test.
- table of sample specific quantifications.
- volcano plots.
- PCA plots computed on the samples.
- correlation matrix plots computed on the samples.

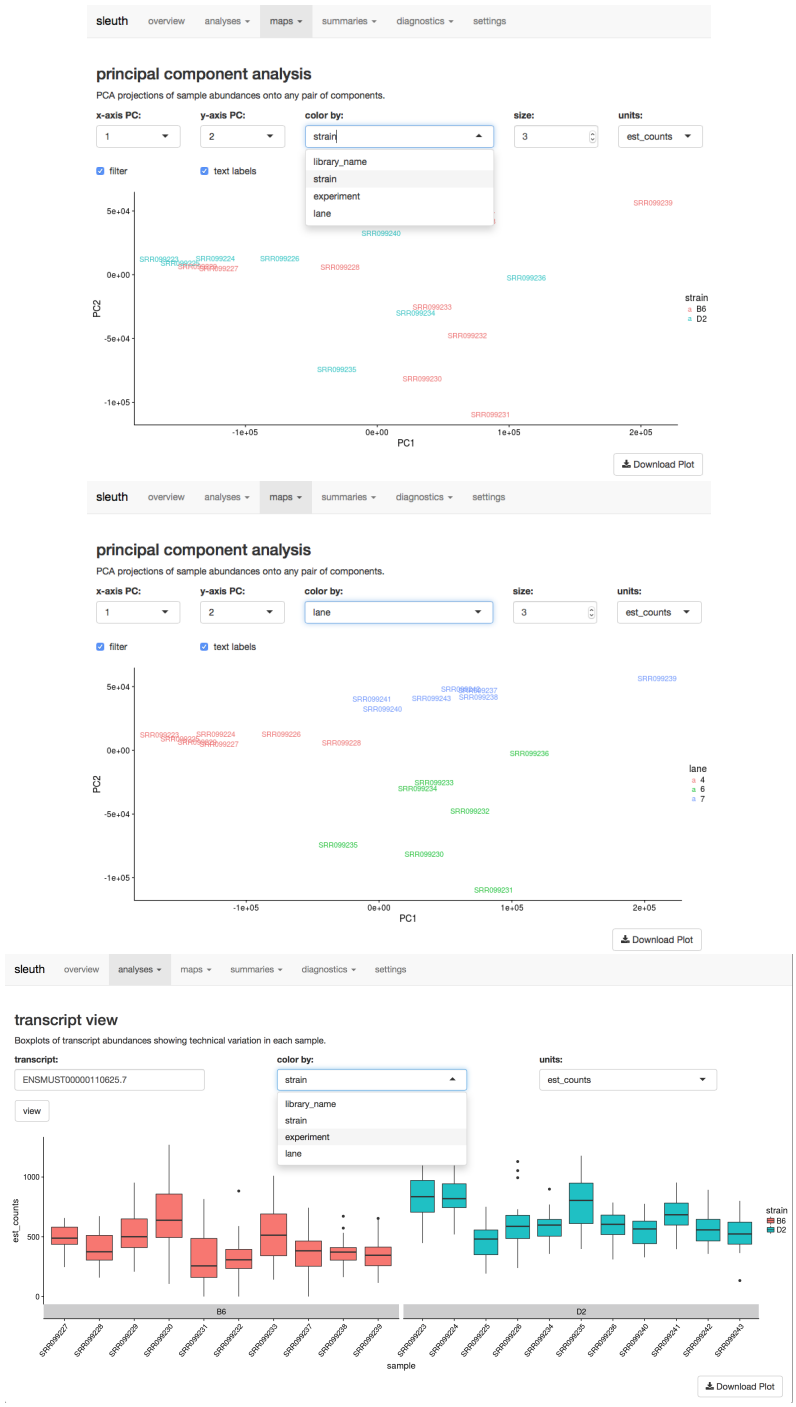
Nature Methods: doi:10.1038/nmeth.4324

- density plots per sample.
- display of the design matrices.
- plots of sample specific fragment length distribution.
- table of sample specific variables.
- table of sample specific bias parameters.
- plot of the mean-variance relationship.
- between sample scatter plots.
- Q-Q plots.

Additionally, since we are using Shiny, sleuth objects can be exported and shared with collaborators. These objects will instantly provide the entire analysis performed, including interactive analyses. Alternatively, sleuth objects can be uploaded to a Shiny server where anyone with internet access can automatically view the entire analysis by simply visiting a URL. We have taken this latter approach by providing reanalysis of several publicly available datasets via the Lair [**?**]: https://pachterlab.github.io/lair/

Supplementary Figure SN1 shows some screenshots from a sleuth analysis of the Bottomly data [**?**]. Supplementary Figure 1a shows the principal component analysis of the data set colored by the different conditions. One can see that the first two principal components do not segregate the data by experimental condition (mouse strain). Supplementary Figure 1b shows how one can use the drop-down menu to change the coloring, revealing that the first two principal components seem to explain some of the variation due to the batch. In addition, there are many other features assisting in exploring the data, such as the ability to view, sort and search the table of differential expression results. For example, sorting by the inferential variability and then by largest p-values, we find transcript *Ppip5k1-004* (ENSMUST00000110625), which is not reported as differentially expressed by sleuth, but is reported as differentially expressed by both voom and DESeq2. This is likely due to the high inferential variability which is not being properly assessed and adjusted for by those programs. The transcript name can be pasted into the "transcript view" window and the distribution of inferential variability can be explored with boxplots describing the variability within each sample (Supplementary Figure SN1c).

Nature Methods: doi:10.1038/nmeth.4324



Supplementary Figure SN1: Interactive sleuth live Shiny interface on the complete Bottomly data set. (a) PCA plot colored by strain shows that the strain does not explain much of the variance in the first two principal components. The coloring can be changed immediately by drop-down as shown in (b) which indicates that there are possible lane effects. (c) Sample specific bootstraps for transcript *Ppip5k1-004* (ENSMUST00000110625), which does not show differential expression by sleuth (FDR 0.569), but shows differential expression by voom and DESeq2 (FDR 0.013 and FDR 0.004, respectively). A possible explanation for this is that the inferential variance is quite high.

Nature Methods: doi:10.1038/nmeth.4324